# Statistics for Data Science: Week 12

Myrto Limnios and Rajita Chandak

Institute of Mathematics – EPFL

rajita.chandak@epfl.ch, myrto.limnios@epfl.ch

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 $\boldsymbol{X} = (\mathbf{1}\ \boldsymbol{W})$, $\beta = (\beta_0\ \boldsymbol{\gamma})^\top$
- 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 $\beta_j$ 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.

- Since $\boldsymbol{Z}_j \perp \mathbf{1}$ for all, $j$, we can estimate $\beta_0$ and $\gamma$ by two separate regressions (orthogonality).
- Least squares estimators based on the *standardized design matrix* 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
$\hookrightarrow$ $\lambda$ called *ridge parameter*.

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

$\hookrightarrow$ 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

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

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!

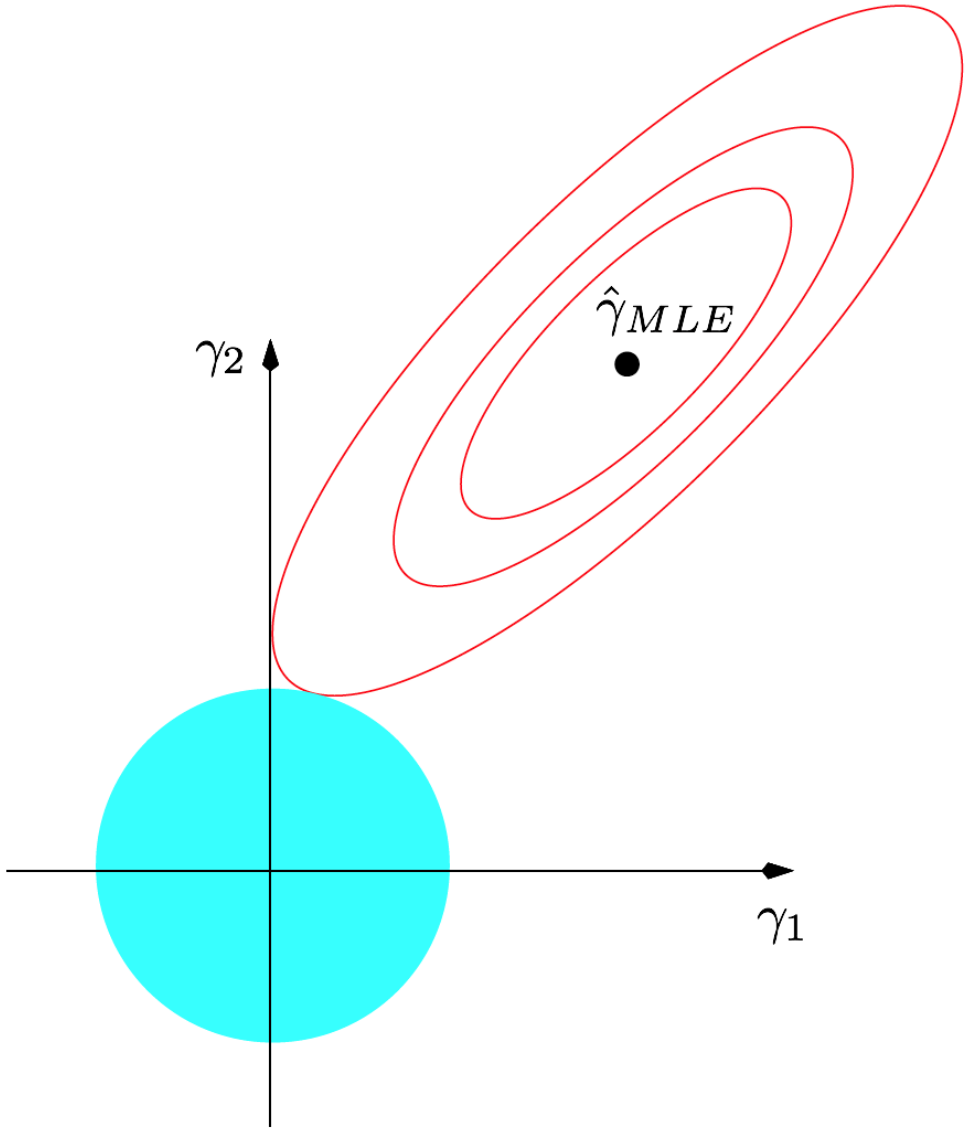



Figure: $L^2$ Shrinkage [Ridge Regression]

## Theorem

*Let $\boldsymbol{Z}_{n\times q}$ be a matrix of rank $r \leq q$ with centred column vectors of unit norm. Given $\lambda > 0$, the unique minimiser of*

$$Q(\alpha, \boldsymbol{\xi}) = \|\boldsymbol{Y} - \alpha\mathbf{1} - \boldsymbol{Z}\boldsymbol{\xi}\|_2^2 + \lambda\|\boldsymbol{\xi}\|_2^2$$

*is*

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

## Proof.

Write

$$\boldsymbol{Y} = \underbrace{(\boldsymbol{Y} - \bar{Y}\mathbf{1})}_{=\boldsymbol{Y}^*\in\mathcal{M}^\perp(\mathbf{1})} + \underbrace{\bar{Y}\mathbf{1}}_{\in\mathcal{M}(\mathbf{1})}$$

Note also that by assumption $\mathbf{1} \in \mathcal{M}^\perp(\boldsymbol{Z})$. Therefore by Pythagoras' theorem

$$\|\boldsymbol{Y} - \hat{\beta}_0\mathbf{1} - \boldsymbol{Z}\hat{\boldsymbol{\gamma}}\|_2^2 = \|\underbrace{(\bar{Y} - \hat{\beta}_0)\mathbf{1}}_{\in\mathcal{M}(\mathbf{1})} + \underbrace{(\boldsymbol{Y}^* - \boldsymbol{Z}\hat{\boldsymbol{\gamma}})}_{\in\mathcal{M}(\boldsymbol{Z})}\|_2^2 = \|(\bar{Y} - \hat{\beta}_0)\mathbf{1}\|_2^2 + \|(\boldsymbol{Y}^* - \boldsymbol{Z}\hat{\boldsymbol{\gamma}})\|_2^2.$$

Therefore, $\min_{\alpha,\boldsymbol{\xi}} Q(\alpha,\boldsymbol{\xi}) = \min_{\alpha} \|(\bar{Y}-\alpha)\mathbf{1}\|_2^2 + \min_{\boldsymbol{\xi}} \left\{ \|(\boldsymbol{Y}^* - \boldsymbol{Z}\boldsymbol{\xi})\|_2^2 + \lambda\|\boldsymbol{\xi}\|_2^2 \right\}$

Clearly, $\arg\min_\alpha \|(\bar{Y}-\alpha)\mathbf{1}\|_2^2 = \bar{Y}$ while the second component can be written

$$\min_{\boldsymbol{\xi}\in\mathbb{R}^q} \left\| \begin{pmatrix} \boldsymbol{Z} \\ \sqrt{\lambda} I_{q\times q} \end{pmatrix} \boldsymbol{\xi} - \begin{pmatrix} \boldsymbol{Y}^* \\ \mathbf{0}_{q\times 1} \end{pmatrix} \right\|_2^2$$

using block notation. This is the usual least squares problem with solution

$$\left[ (\boldsymbol{Z}^\top, \sqrt{\lambda}\boldsymbol{I}_{q\times q}) \begin{pmatrix} \boldsymbol{Z} \\ \sqrt{\lambda}\boldsymbol{I}_{q\times q} \end{pmatrix} \right]^{-1} (\boldsymbol{Z}^\top, \sqrt{\lambda}\boldsymbol{I}_{q\times q}) \begin{pmatrix} \boldsymbol{Y}^* \\ \mathbf{0}_{q\times 1} \end{pmatrix} = (\boldsymbol{Z}^\top\boldsymbol{Z} + \lambda\boldsymbol{I})^{-1}\boldsymbol{Z}^\top\boldsymbol{Y}^*$$

Note that $\boldsymbol{Z}^\top\boldsymbol{Z} + \lambda\boldsymbol{I}$ is indeed invertible. Writing $\boldsymbol{Z}^\top\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, we have

$$\boldsymbol{Z}^\top\boldsymbol{Z} + \lambda\boldsymbol{I} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top + \boldsymbol{U}(\lambda\boldsymbol{I}_{q\times q})\boldsymbol{U}^\top = \boldsymbol{U}(\boldsymbol{\Lambda} + \lambda\boldsymbol{I}_{q\times q})\boldsymbol{U}^\top$$

and $\boldsymbol{\Lambda} = \text{diag}\{\underbrace{\lambda_1,\ldots,\lambda_r}_{>0}, \underbrace{\lambda_{r+1},\ldots,\lambda_q}_{=0}\}$ ($\boldsymbol{Z}^\top\boldsymbol{Z} \succeq 0$ & $\text{rank}(\boldsymbol{Z}^\top\boldsymbol{Z}) = \text{rank}(\boldsymbol{Z})$).

To complete the proof, observe that $\boldsymbol{Z}^\top\boldsymbol{Y}^* = \boldsymbol{Z}^\top\boldsymbol{Y} - \bar{Y}\boldsymbol{Z}^\top\mathbf{1} = \boldsymbol{Z}^\top\boldsymbol{Y}$. □

Note that if the SVD of $\boldsymbol{Z}$ is $\boldsymbol{Z} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, last steps of previous proof may be used to show that

$$\hat{\gamma} = \sum_{j=1}^{q} \frac{\lambda_j}{\lambda_j^2 + \lambda} (\boldsymbol{V}_j^\top \boldsymbol{Y}) \boldsymbol{U}_j,$$

where the $\boldsymbol{V}_j$s and $\boldsymbol{U}_j$s are the columns of $\boldsymbol{V}$ and $\boldsymbol{U}$, respectively.

Compare this to the ordinary least squares solution, when $\lambda = 0$:

$$\hat{\gamma} = \sum_{j=1}^{q} \frac{1}{\lambda_j} (\boldsymbol{V}_j^\top \boldsymbol{Y}) \boldsymbol{U}_j,$$

which is not even defined if $\boldsymbol{Z}$ is of reduced rank.

Role of $\lambda$ is to reduce the size of $1/\lambda_j$ when $\lambda_j$ becomes very small.

## Proposition

*Let $\hat{\gamma}$ be the ridge regression estimator of $\gamma$. Then*

$$\text{bias}(\hat{\gamma}, \gamma) = -\lambda \left( \boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I}_{q \times q} \right)^{-1} \gamma$$

*and*

$$\text{cov}(\hat{\gamma}) = \sigma^2 (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1} \boldsymbol{Z}^\top \boldsymbol{Z} (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1}.$$

## Proof.

Since $\mathbb{E}(\hat{\gamma}) = (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1} \boldsymbol{Z}^\top \mathbb{E}(\boldsymbol{Y}) = (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1} \boldsymbol{Z}^\top \boldsymbol{Z} \gamma$, the bias is

$$\begin{aligned}
\text{bias}(\hat{\gamma}, \gamma) &= \mathbb{E}(\hat{\gamma}) - \gamma = \{ (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1} \boldsymbol{Z}^\top \boldsymbol{Z} - \boldsymbol{I} \} \gamma \\
&= \left\{ \left( \frac{1}{\lambda} \boldsymbol{Z}^\top \boldsymbol{Z} + \boldsymbol{I} \right)^{-1} \left( \frac{1}{\lambda} \boldsymbol{Z}^\top \boldsymbol{Z} + \boldsymbol{I} - \boldsymbol{I} \right) - \boldsymbol{I} \right\} \gamma \\
&= \left\{ \boldsymbol{I} - \left( \frac{1}{\lambda} \boldsymbol{Z}^\top \boldsymbol{Z} + \boldsymbol{I} \right)^{-1} - \boldsymbol{I} \right\} \gamma = - \left( \frac{1}{\lambda} \boldsymbol{Z}^\top \boldsymbol{Z} + \boldsymbol{I} \right)^{-1} \gamma.
\end{aligned}$$

The covariance term is straightforward. □

## Corollary (Domination over Least Squares)

*Assume that* $\operatorname{rank}(\boldsymbol{Z}_{n\times q}) = q$ *and let*

$$\tilde{\gamma} = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{Y} \qquad \& \qquad \hat{\gamma}_\lambda = (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1}\boldsymbol{Z}^\top \boldsymbol{Y}$$

*be the least squares estimator and ridge estimator, respectively. Then,*

$$\mathbb{E}\left\{(\tilde{\gamma}-\gamma)(\tilde{\gamma}-\gamma)^\top\right\} - \mathbb{E}\left\{(\hat{\gamma}_\lambda-\gamma)(\hat{\gamma}_\lambda-\gamma)^\top\right\} \succeq 0$$

*for all* $\lambda \in (0, 2\sigma^2/\|\gamma\|^2)$.

Ridge estimator uniformly better than least squares! How can this be?
*(What happened to Gauss-Markov?)*

- Gauss-Markov only covers unbiased estimators – but ridge estimator biased.
- A bit of bias can improve the MSE by reducing variance!
- Also, there is a catch! The "right" range for $\lambda$ depends on unknowns.
- Choosing a good $\lambda$ is all about balancing bias and variance.

## Proof.

From our bias/variance calculations on the ridge estimator, we have

$$\mathbb{E}\left\{(\tilde{\gamma}-\gamma)(\tilde{\gamma}-\gamma)^\top\right\} - \mathbb{E}\left\{(\hat{\gamma}_\lambda-\gamma)(\hat{\gamma}_\lambda-\gamma)^\top\right\} =$$

$$\sigma^2(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1} - (\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1}\sigma^2\boldsymbol{Z}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1} - \lambda^2\left(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I}\right)^{-1}\gamma\gamma^\top\left(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I}\right)^{-1}$$

$$= \lambda(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1}\Big(\sigma^2(2\boldsymbol{I}+\lambda(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}) - \lambda\gamma\gamma^\top\Big)(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1}.$$

To go from 2nd to 3rd line, we wrote

$$\sigma^2(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1} = \sigma^2(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1}(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1}$$

$$= (\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1}(\sigma^2\boldsymbol{Z}^\top\boldsymbol{Z}+2\sigma^2\lambda\boldsymbol{I}+\sigma^2\lambda^2(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1})(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1}$$

and did the tedious (but straighforward) algebra. The final term can be made positive definite if

$$2\sigma^2\boldsymbol{I}+\sigma^2\lambda(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}-\lambda\gamma\gamma^\top \succeq 0.$$

Noting that we can always write

$$\boldsymbol{I} = \frac{\gamma\gamma^\top}{\|\gamma\|^2} + \sum_{j=1}^{q-1}\boldsymbol{\theta}_j\boldsymbol{\theta}_j^\top$$

for $\{\gamma/\|\gamma\|, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{q-1}\}$ an orthonormal basis of $\mathbb{R}^q$ we see that $\lambda \in (0, 2\sigma^2/\|\gamma\|^2)$ suffices for positive definiteness to hold true.

□

Role of $\lambda$: Regulates Bias–Variance tradeoff

- $\lambda \uparrow$ decreases variance (collinearity) but increases bias
- $\lambda \downarrow$ decreases bias but variance inflated if collinearity exists

Recall:

$$\mathbb{E}\|\hat{\gamma} - \gamma\|^2 = \underbrace{\mathbb{E}\|\mathbb{E}\hat{\gamma} - \gamma\|^2}_{\text{bias}^2} + \underbrace{\mathbb{E}\|\hat{\gamma} - \mathbb{E}\hat{\gamma}\|^2}_{\text{variance}=\textit{trace}[\textit{cov}(\hat{\gamma})]} + \underbrace{2(\mathbb{E}\hat{\gamma} - \gamma)^\top \mathbb{E}[\hat{\gamma} - \mathbb{E}\hat{\gamma}]}_{=0}$$

Writing $\boldsymbol{Z}^\top \boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ $\text{trace}\{\text{cov}(\hat{\gamma})\} = \sum_{j=1}^{q} \frac{\lambda_i}{\lambda_i^2+\lambda}\sigma^2$

So choose $\lambda$ so as to optimally to balance bias/variance

Use cross validation!

Motivated from Ridge Regression formulation can consider:

$$\text{min!} \qquad \|\boldsymbol{Y} - \beta_0\mathbf{1} - \boldsymbol{Z}\boldsymbol{\gamma}\|_2^2 \quad \text{subject to} \quad \sum_{j=1}^{p-1} |\gamma_j| = \|\boldsymbol{\gamma}\|_1 \leq r(\lambda)$$

$$\Longleftrightarrow$$

$$\text{min!} \qquad \|\boldsymbol{Y} - \beta_0\mathbf{1} - \boldsymbol{Z}\boldsymbol{\gamma}\|_2^2 + \lambda\|\boldsymbol{\gamma}\|_1.$$

Shrinks coefficient *size* by different version of *magnitude*.

- Resulting estimator non–linear in $\boldsymbol{Y}$
- No explicit form available (unless $\boldsymbol{Z}^\top\boldsymbol{Z} = \boldsymbol{I}$), needs quadratic programming algorithm

Why choose a different type of norm?
$L^1$ penalty (almost) produces a "continuous" model selection!

When the explanatory variables are orthogonal (i.e. $\boldsymbol{Z}^\top \boldsymbol{Z} = \boldsymbol{I}$), then the LASSO[1] exactly performs model selection via thresholding:

## Theorem (Orthogonal Design $\leftrightarrow$ Model Selection)

*Consider the linear model*

$$\underset{n\times 1}{\boldsymbol{Y}} = \beta_0 \underset{1\times 1}{\mathbf{1}}_{n\times 1} + \underset{n\times(p-1)}{\boldsymbol{Z}} \underset{(p-1)\times 1}{\gamma} + \underset{n\times 1}{\varepsilon}$$

*where* $\boldsymbol{Z}^\top \mathbf{1} = 0$ *and* $\boldsymbol{Z}^\top \boldsymbol{Z} = I$. *Let* $\hat{\gamma}$ *be the least squares estimator of* $\gamma$,

$$\hat{\gamma} = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{Y} = \boldsymbol{Z}^\top \boldsymbol{Y}.$$

*Then, the unique solution to the LASSO problem*

$$\min_{\beta_0 \in \mathbb{R}, \gamma \in \mathbb{R}^{p-1}} \left\{ \|\boldsymbol{Y} - \beta_0 \mathbf{1} - \boldsymbol{Z}\gamma\|_2^2 + \lambda \|\gamma\|_1 \right\}$$

*is given by* $(\hat{\beta}_0, \check{\gamma}) = (\beta_0, \check{\gamma}_1, \ldots, \check{\gamma}_{p-1})$, *defined as*

$$\hat{\beta}_0 = \bar{Y} \qquad \& \qquad \check{\gamma}_i = \operatorname{sgn}(\hat{\gamma}_i) \left( |\hat{\gamma}_i| - \frac{\lambda}{2} \right)_+, \quad i = 1, ..., p-1.$$

[1] Least Absolute Shrinkage and Selection Operator.

## Proof $(*)$.

Note that since $Z^\top \mathbf{1} = 0$ and since $\beta_0$ does not appear in the $L^1$ penalty, we have

$$\hat{\beta}_0 = (\mathbf{1}^\top \mathbf{1})^{-1} \mathbf{1} Y = \bar{Y}.$$

Thus, the LASSO problem reduces to

$$\min_{\beta_0 \in \mathbb{R}, \boldsymbol{\gamma} \in \mathbb{R}^{p-1}} \left\{ \| \boldsymbol{Y} - \beta_0 \mathbf{1} - Z\boldsymbol{\gamma} \|_2^2 + \lambda \|\boldsymbol{\gamma}\|_1 \right\} = \min_{\boldsymbol{\gamma} \in \mathbb{R}^{p-1}} \left\{ \| \boldsymbol{u} - \boldsymbol{Z}\boldsymbol{\gamma} \|_2^2 + \lambda \|\boldsymbol{\gamma}\|_1 \right\}.$$

where $\boldsymbol{u} = \boldsymbol{Y} - \bar{Y}\mathbf{1}$ for tidiness. Expanding $\| \boldsymbol{u} - \boldsymbol{Z}\boldsymbol{\gamma} \|_2^2$ gives

$$\boldsymbol{u}^\top \boldsymbol{u} - 2\boldsymbol{u}^\top \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\gamma}^\top \underbrace{(\boldsymbol{Z}^\top \boldsymbol{Z})}_{=I} \boldsymbol{\gamma} = \boldsymbol{u}^\top \boldsymbol{u} - 2\underbrace{\boldsymbol{Y}^\top \boldsymbol{Z}}_{=\hat{\boldsymbol{\gamma}}^\top} \boldsymbol{\gamma} + 2\bar{Y} \underbrace{\mathbf{1}^\top \boldsymbol{Z}}_{=0} \boldsymbol{\gamma} + \boldsymbol{\gamma}^\top \boldsymbol{\gamma}$$

Since $\boldsymbol{u}^\top \boldsymbol{u}$ does not depend on $\boldsymbol{\gamma}$, we see that the LASSO objective function is

$$-2\hat{\boldsymbol{\gamma}}^\top \boldsymbol{\gamma} + \|\boldsymbol{\gamma}\|_2^2 + \lambda \|\boldsymbol{\gamma}\|_1.$$

Clearly, this has the same minimizer if multiplied across by $1/2$, i.e.

$$-\hat{\boldsymbol{\gamma}}^\top \boldsymbol{\gamma} + \tfrac{1}{2}\|\boldsymbol{\gamma}\|_2^2 + \tfrac{1}{2}\lambda \|\boldsymbol{\gamma}\|_1 = \textstyle\sum_{i=1}^{p-1} \left( -\hat{\gamma}_i \gamma_i + \frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}|\gamma_i| \right).$$

Notice that we now have a sum of $p-1$ objective functions, each depending only on one $\gamma_i$. We can thus optimise each separately. That is, for any given $i \leq p-1$, we must minimise

$$-\hat{\gamma}_i \gamma_i + \frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}|\gamma_i|.$$

We distinguish 3 cases:

1. Case $\hat{\gamma}_i = 0$. In this case, the objective function becomes $\frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}|\gamma_i|$ and it is clear that it is minimised when $\gamma_i = 0$. So in this case $\check{\gamma}_i = 0$.
2. Case $\hat{\gamma}_i > 0$. In this case, the objective function $-\hat{\gamma}_i \gamma_i + \frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}|\gamma_i|$ is minimised somewhere in the range $\gamma_i \in [0, \infty)$ because the term $-\hat{\gamma}_i \gamma_i$ is negative there (and all other terms are positive). But when $\gamma_i \geq 0$, the objective function becomes

$$-\hat{\gamma}_i \gamma_i + \frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}\gamma_i = \left(\frac{\lambda}{2} - \hat{\gamma}_i\right)\gamma_i + \frac{1}{2}\gamma_i^2.$$

   If $\frac{\lambda}{2} - \hat{\gamma}_i \geq 0$, then the minimum over $\gamma_i \in [0, \infty)$ is clearly at $\gamma_i = 0$. Otherwise, when $\frac{\lambda}{2} - \hat{\gamma}_i < 0$, we differentiate and find the minimum at $\gamma_i = \hat{\gamma}_i - \lambda/2 > 0$. In summary, $\check{\gamma}_i = (\hat{\gamma}_i - \lambda/2)_+ = \mathrm{sgn}(\hat{\gamma}_i)(|\hat{\gamma}_i| - \lambda/2)_+$.

3. Case $\hat{\gamma}_i < 0$. In this case, the objective function $-\hat{\gamma}_i\gamma_i + \frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}|\gamma_i|$ is minimised somewhere in the range $\gamma_i \in (-\infty, 0]$ because the term $-\hat{\gamma}_i\gamma_i$ is negative there (and all other terms are positive). But when $\gamma_i \leq 0$, the objective function becomes

$$-\hat{\gamma}_i\gamma_i + \frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}(-\gamma_i) = \left(\frac{\lambda}{2} + \hat{\gamma}_i\right)(-\gamma_i) + \frac{1}{2}\gamma_i^2 = \left(\frac{\lambda}{2} - |\hat{\gamma}_i|\right)(-\gamma_i) + \frac{1}{2}\gamma_i^2.$$

If $\frac{\lambda}{2} - |\hat{\gamma}_i| \geq 0$, then the minimum over $\gamma_i \in (-\infty, 0]$ is clearly at $\gamma_i = 0$, since $-\gamma_i$ ranges over $[0, \infty)$. Otherwise, when $\frac{\lambda}{2} - |\hat{\gamma}_i| < 0$, we differentiate and find the minimum at $\gamma_i = -|\hat{\gamma}_i| + \lambda/2 < 0$, which we may re-write as:

$$-|\hat{\gamma}_i| + \lambda/2 = -\left(|\hat{\gamma}_i| - \lambda/2\right) = \operatorname{sgn}(\hat{\gamma}_i)\left(|\hat{\gamma}_i| - \lambda/2\right).$$

In summary, $\check{\gamma}_i = \operatorname{sgn}(\hat{\gamma}_i)(|\hat{\gamma}_i| - \lambda/2)_+$.

The proof is now complete, as we can see that all three cases yield

$$\check{\gamma}_i = \operatorname{sgn}(\hat{\gamma}_i)\left(|\hat{\gamma}_i| - \frac{\lambda}{2}\right)_+, \quad i = 1, ..., p-1.$$

□

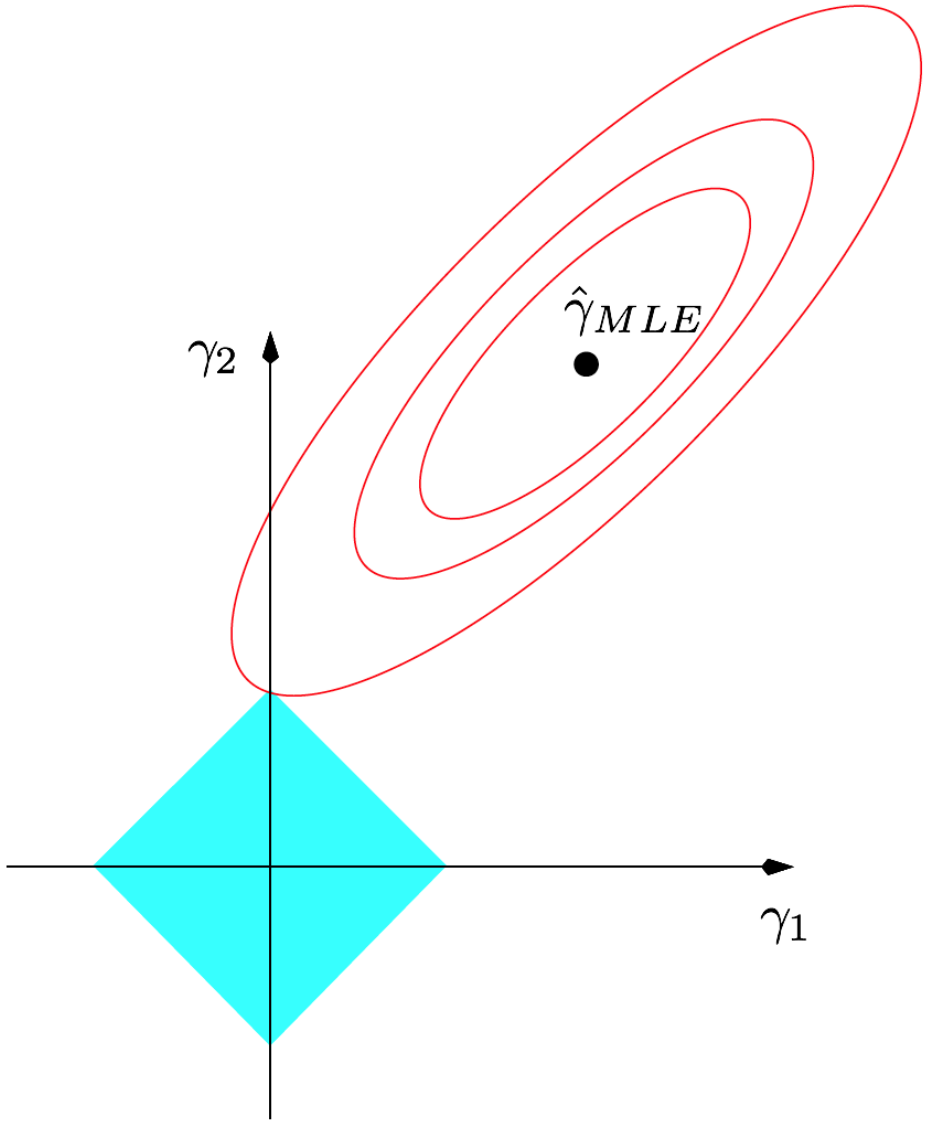



Figure: $L^1$ Shrinkage (the LASSO)

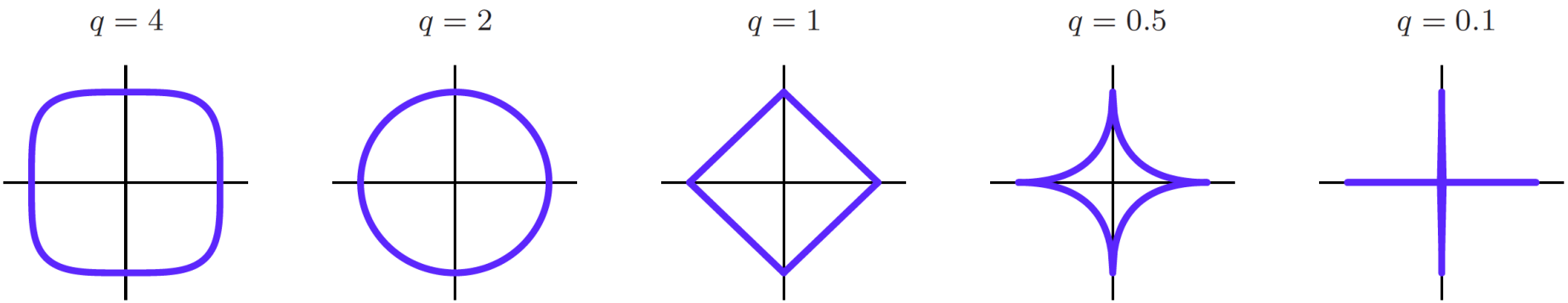
Intuition: $L_1$ norm induces "sharp" balls!

- Balls more concentrated around the axes
- Induces model selection by regulating the LASSO (through $\lambda$)

Extreme case: $L^0$ "Norm", gives best subset selection!

$$\|\boldsymbol{\gamma}\|_0 = \sum_{j=1}^{p-1} |\gamma_j|^0 = \sum_{j=1}^{p-1} \mathbf{1}_{\{\gamma_j \neq 0\}} = \#\{j : \gamma_j \neq 0\}$$

Generally: $\|\boldsymbol{\gamma}\|_q^q = \sum_{j=1}^{p-1} |\gamma_j|^q$, sharp balls for $0 < q \leq 1$




But $L^1$ gives sharpest convex ball among these.

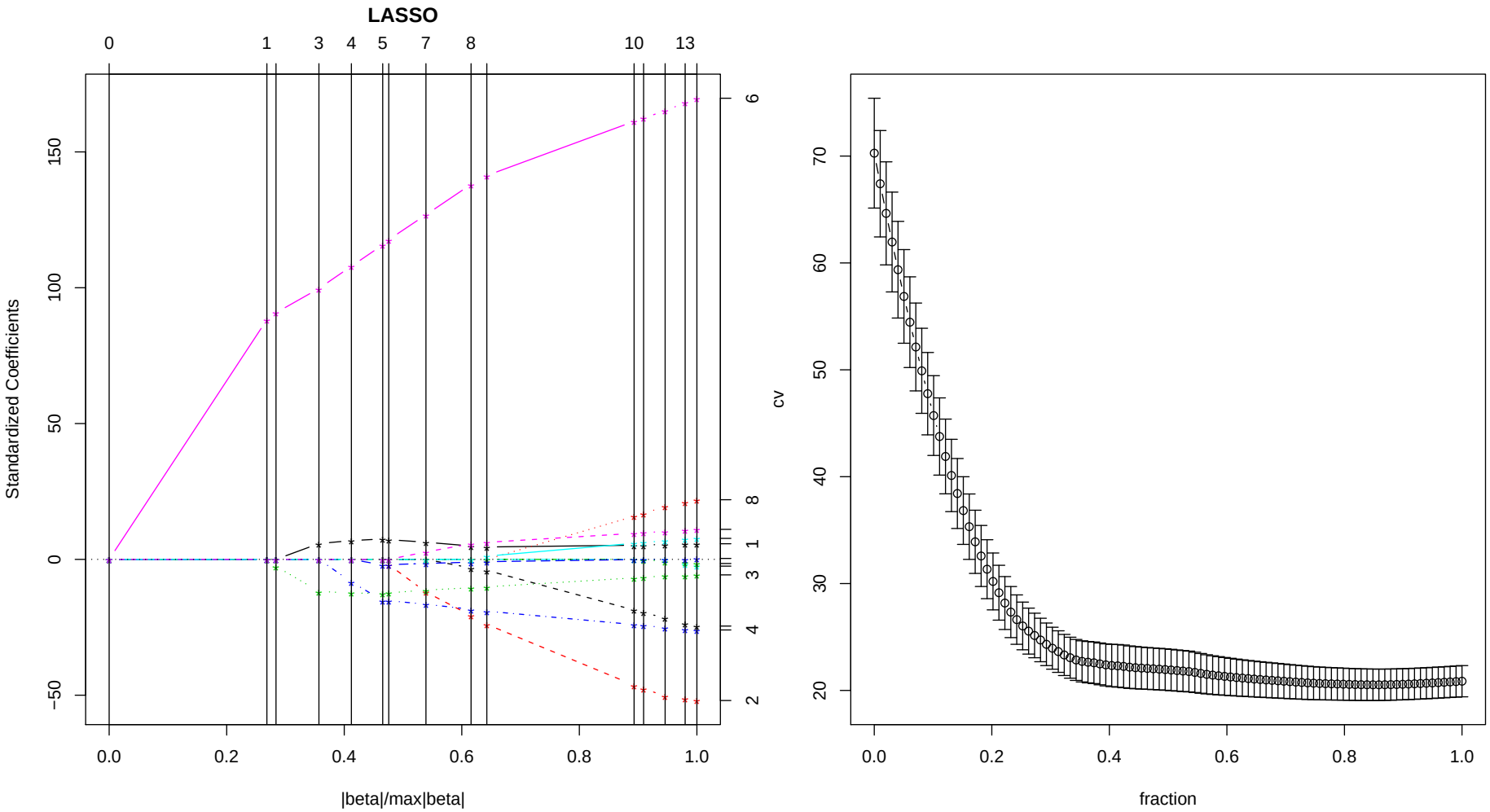



Figure: LASSO and CV for different values of $r(\lambda)/\|\hat{\gamma}\|_1$ for the bodyfat data (LARS algorithm), on the right: fraction of the final $L^1$ norm

- Regularization achieves smaller variance at the price of non-zero bias
- Need to find the best trade-off between variance/bias
- The larger the parameter $\lambda$, the greater the shrink (so different values of $\lambda$ imply different estimates for $\gamma$)
- Ridge regression is a convex objective risk, and works well when there is a subset of coefficients that are small
- LASSO is used as *feature selection* method, as it yields estimates of parameters to be exactly equal to 0
- The tuning parameter $\lambda$ tailors the strength of the penalization
- Choosing numerically the tuning parameter is often by $k$-fold cross validation (and NOT based on training error): we want to avoid over-/under-fitting

# Generalised Linear Models

Back to the big picture:

$$Y \text{ (random output)} \overset{\text{whose law is influenced by}}{\longleftarrow} x \text{ (non-random input)}$$

General formulation:

$$Y_i \overset{\text{independent}}{\sim} \text{Distribution}\{\underbrace{g(x_i)}_{=\theta_i}\}, \qquad i = 1, ..., n.$$

| **Distribution / Function** $g$ | $g(\boldsymbol{x}_i^\top) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ | $g$ nonparametric |
|---|---|---|
| Gaussian | Linear Regression ✓ | Smoothing |
| Exponential Family | GLM ← | GAM |

Generalised Linear Models: regression with exponential family responses!

## GLM for $Y_1, ..., Y_n$

Response vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ has independent entries with joint law

$$f(\boldsymbol{y}; \boldsymbol{\phi}) = \prod_{i=1}^{n} \exp\{\phi_i y_i - \gamma(\phi_i) + S(y_i)\} = \exp\left\{\boldsymbol{\phi}^\top \boldsymbol{y} - \sum_{i=1}^{n} \gamma(\phi_i) + \sum_{i=1}^{n} S(y_i)\right\},$$

where $\boldsymbol{\phi} \in \Phi \subseteq \mathbb{R}^n$ is the natural parameter with $\Phi$ open. The parameter varies as a function of the covariates via

$$\boldsymbol{\phi} = \boldsymbol{X}_n \boldsymbol{\beta},$$

for $\boldsymbol{X}_n$ the $n \times p$ covariate matrix of rank $p$ and $\boldsymbol{\beta}$ a $p$-dimensional parameter.

In our general notation

$$Y_i \overset{\text{independent}}{\sim} f(y_i; \phi_i) = \exp\{\underbrace{(\boldsymbol{x}_i^\top \boldsymbol{\beta})}_{=\phi_i} y_i - \gamma(\underbrace{\boldsymbol{x}_i^\top \boldsymbol{\beta}}_{=\phi_i}) + S(y_i)\}, \qquad i = 1, ..., n.$$

where the row vector $\boldsymbol{x}_i^\top$ is the $i$th row of $\boldsymbol{X}_n$.

Comments:

- Notice that the sufficient statistic for each marginal distribution $f(y;\phi_i)$ was taken to be the identity $T(Y) = Y$.

- This does not incur any loss of generality for two reasons:

  1. In the three main GLM of interest (Gaussian, Bernoulli, and Poisson) the natural statistic is for $f(y;\phi_i)$ is indeed the identity.
  2. More generally, since we only observe a single observation from each $f(y;\phi_i)$, if the natural statistic were $T(Y_i) \neq Y_i$, we could re-define the response to just be $T_i = T(Y_i)$. The sampling distribution of $T_i$ can be shown to also be a one-parameter exponential family with the same natural parameter.

- Recall from our sampling theory results:

  - $\mu_i = \mathbb{E}[Y_i] = \dfrac{\partial}{\partial \phi_i}\gamma(\phi_i)$
  - $\mathrm{var}[Y_i] = \dfrac{\partial^2}{\partial \phi_i^2}\gamma(\phi_i)$
  - So if $\gamma$ is invertible (which it is for the three main examples), the variance can be written as a function of the mean:

$$\mathrm{var}[Y_i] = \gamma''([\gamma']^{-1}(\mu_i)) = V(\mu_i).$$

## Interpretation of $\phi_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ ?

- In key cases $\phi_i$ is directly interpretable. If not, can switch perspective using the mean $\mu_i$ as defining parameter, connected to the linear predictor $\boldsymbol{x}_i^\top \boldsymbol{\beta}$ via

$$[\gamma']^{-1}(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} = \phi_i$$

- The function $[\gamma']^{-1}(\cdot)$ is called the natural link function.
- Instead of $[\gamma']^{-1}$ can use other link functions $g(\cdot)$ and postulate

$$g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

  This will also yield a GLM, but now the natural parameter will not be equal to the linear predictor but to some function $u(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ of it.
- In summary, the nomenclature is:
  - $g(\cdot)$ is the link function
  - $h = g^{-1}$ is the inverse link function
  - $g(\cdot) = [\gamma']^{-1}(\cdot)$ is the natural link function
- Will focus on natural links but methods/results generalise quite easily.

With natural link, the loglikelihood (up to constants w.r.t. $\boldsymbol{\beta}$) is

$$\ell_n(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{X}_n^\top \boldsymbol{Y} - \sum_{i=1}^{n} \gamma(\boldsymbol{x}_i^\top \boldsymbol{\beta})$$

for $\boldsymbol{x}_i^\top$ the $i$th row of $\boldsymbol{X}_n$. The corresponding $p \times 1$ derivative (*score function*) is

$$\nabla_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}) = \boldsymbol{X}_n^\top \boldsymbol{Y} - \sum_{i=1}^{n} \boldsymbol{x}_i \gamma'(\boldsymbol{x}_i^\top \boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i \left(Y_i - \mu_i\right) = \boldsymbol{X}_n^\top (\boldsymbol{Y} - \boldsymbol{\mu})$$

with $p \times p$ covariance equaling the information matrix and given by

$$\begin{aligned} \operatorname{cov}\left\{\nabla_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta})\right\} &= \sum_{i=1}^{n} \boldsymbol{x}_i^\top \operatorname{cov}\left\{Y_i - \mu_i\right\} \boldsymbol{x}_i = \boldsymbol{X}_n^\top \boldsymbol{V}(\boldsymbol{\beta}) \boldsymbol{X}_n \\ &= -\nabla^2_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}) = \mathcal{I}_n(\boldsymbol{\beta}), \end{aligned}$$

where $\operatorname{cov}\{\boldsymbol{Y}\} = \boldsymbol{V}(\boldsymbol{\beta}) \succ 0$ is diagonal, with $i$th diagonal element

$$\operatorname{var}\{Y_i\} = \gamma''(\phi_i) = \gamma''(\boldsymbol{x}_i^\top \boldsymbol{\beta}).$$

Thus, if the MLE $\hat{\boldsymbol{\beta}}$ exists it is also unique, and must satisfy:

$$\sum_{i=1}^{n} \boldsymbol{x}_i \left( Y_i - \gamma'(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}) \right) = 0$$

By a first order Taylor expansion of $\gamma'$, we have

$$\gamma'(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}) \approx \gamma'(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) + \boldsymbol{x}_i^\top (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \gamma''(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}})$$

for some guesstimate $\tilde{\boldsymbol{\beta}}$ near $\hat{\boldsymbol{\beta}}$. Plugging into the score equation yields

$$\sum_{i=1}^{n} \boldsymbol{x}_i^\top \left( Y_i - \gamma'(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) + \boldsymbol{x}_i^\top (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \gamma''(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) \right) \approx 0$$

$$\implies \sum_{i=1}^{n} \gamma''(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) \boldsymbol{x}_i^\top (Z_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}) = \boldsymbol{X}_n^\top \boldsymbol{V}(\tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{Z}} - \boldsymbol{X}_n \hat{\boldsymbol{\beta}}) \approx 0$$

where we defined $\tilde{\boldsymbol{Z}} = (\tilde{Z}_1, ..., \tilde{Z}_n)^\top$ to be the adjusted response

$$\tilde{Z}_i = \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}} + \frac{1}{\gamma''(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}})} (Y_i - \gamma'(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}})) \quad \text{so} \quad \tilde{\boldsymbol{Z}} = \boldsymbol{X}_n \tilde{\boldsymbol{\beta}} + \boldsymbol{V}^{-1}(\tilde{\boldsymbol{\beta}})(\boldsymbol{Y} - \boldsymbol{\mu}(\tilde{\boldsymbol{\beta}})).$$

The last expression for the score expression now yields:

$$\hat{\beta} \approx (\boldsymbol{X}_n^\top \boldsymbol{V}(\tilde{\beta})\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^\top \boldsymbol{V}(\tilde{\beta})\tilde{\boldsymbol{Z}}$$

Just a weighted least squares estimate! Where's the catch?

- Weight matrix $\boldsymbol{V}(\tilde{\beta})$ requires specification of an initial guesstimate $\boldsymbol{x}_i^\top \tilde{\beta}$ sufficiently close to $\boldsymbol{x}_i^\top \hat{\beta}$.
- Luckily we can give such a guesstimate by recalling that $\mu_i = \gamma'(\boldsymbol{x}_i^\top \beta)$ so that estimating $\mu_i$ by $Y_i$ yields the guesstimate $\boldsymbol{x}_i^\top \tilde{\beta} \equiv (\gamma')^{-1}(Y_i)$.
- Suggests the following Iteratively Reweighted Least Squares (IRLS)

## IRLS

1. Initialize with $\boldsymbol{x}_i^\top \beta^{(0)} \leftarrow (\gamma')^{-1}(Y_i)$ and $Z_i^{(0)} = \boldsymbol{x}_i^\top \beta^{(0)} + \frac{(Y_i - \gamma'(\boldsymbol{x}_i^\top \beta^{(0)}))}{\gamma''(\boldsymbol{x}_i^\top \beta^{(0)})}$
2. Update with $\beta^{(j+1)} \leftarrow (\boldsymbol{X}_n^\top \boldsymbol{V}(\beta^{(j)})\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^\top \boldsymbol{V}(\beta^{(j)})\boldsymbol{Z}^{(j)}$

- Equivalent to Newton-Raphson iteration.
- Not always guaranteed to converge.

Heuristics:

- Suppose we had started iteration at true $\beta$ and stopped at first iterate:

$$\hat{\beta} = (\boldsymbol{X}_n^\top \boldsymbol{V}(\beta)\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^\top \boldsymbol{V}(\beta)\boldsymbol{Z}$$

  where

$$Z_i = \boldsymbol{x}_i^\top \beta + \frac{1}{\gamma''(\boldsymbol{x}_i^\top \beta)}(Y_i - \gamma'(\boldsymbol{x}_i^\top \beta)) \quad \text{so} \quad \boldsymbol{Z} = \boldsymbol{X}_n\beta + \boldsymbol{V}^{-1}(\beta)(\boldsymbol{Y} - \boldsymbol{\mu}(\beta))$$

- This would give us,

$$\hat{\beta} = \beta + (\boldsymbol{X}_n^\top \boldsymbol{V}(\beta)\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^\top (\boldsymbol{Y} - \boldsymbol{\mu}(\beta))$$

- So we would expect $\mathbb{E}[\hat{\beta}] = \beta$ and $\mathrm{cov}[\hat{\beta}] = (\boldsymbol{X}_n^\top \boldsymbol{V}^{-1}(\beta)\boldsymbol{X}_n)^{-1} = \mathcal{I}_n^{-1}(\beta)$.
- And we would conjecture a Gaussian limiting law (paralleling the IID setting)

Under conditions, this is indeed what we obtain.

Result stated in form valid for more general (sufficiently regular) link functions.

## Theorem (Asymptotic Normality of MLE in GLM)

*In the same context and notation as before, assume that:*

(C1) $\boldsymbol{\beta} \in B$ *for* $B$ *an open convex subset of* $\mathbb{R}^p$.

(C2) *The* $p \times p$ *matrix* $\boldsymbol{X}_n^\top \boldsymbol{X}_n$ *is of full rank for all* $n$.

(C3) *The information diverges, i.e.* $\lambda_{min}\Big(\mathcal{I}_n(\boldsymbol{\beta})\Big) \to \infty$ *as* $n \to \infty$ *for* $\lambda_{\min}(\cdot)$ *the smallest eigenvalue.*

(C4) *Given any parameter* $\boldsymbol{\beta} \in \mathbb{R}^p$ *it holds that*

$$\sup_{\boldsymbol{\alpha} \in N_\delta(\beta)} \left\| \mathcal{I}_n^{-1/2}(\boldsymbol{\beta}) \mathcal{I}_n^{1/2}(\boldsymbol{\alpha}) - \boldsymbol{I}_{p \times p} \right\| \to 0$$

$\forall\, \delta > 0$, *where* $N_\delta(\beta) = \left\{ \boldsymbol{\alpha} \in \mathbb{R}^p : (\boldsymbol{\alpha} - \beta)^\top \mathcal{I}_n(\boldsymbol{\beta})(\boldsymbol{\alpha} - \beta) \leq \delta \right\}$.

*Then, as* $n \to \infty$, *provided it exists, the MLE* $\hat{\boldsymbol{\beta}}_n$ *of* $\boldsymbol{\beta}_0$ *is unique & satisfies*

$$\mathcal{I}_n^{1/2}(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, \boldsymbol{I}_{p \times p}).$$

- Recall that under canonical link $\mathcal{I}_n(\boldsymbol{\beta}) = \boldsymbol{X}_n^\top \boldsymbol{V}(\boldsymbol{\beta}) \boldsymbol{X}_n$.

Comments on conditions:

- (C1) implies that $\boldsymbol{X}_n\beta$ ranges over an open set, and so $\gamma$ is infinitely differentiable, and our exponential family possesses all moments.
- (C2) is as with our linear model, and essentially means that our covariates should not be perfectly correlated.
- (C3) is similar to the "balanced design" assumption we had for the asymptotics of a non-Gaussian linear model.
- (C1-C3) are up to us: they depend on the design matrix $\boldsymbol{X}_n$, which in principle is for us to choose.
- (C4) asks that the (root) information matrix converge uniformly on compact ellipsoids centred at the true parameter.

Comments on conclusion:

- Can also be read as saying that for $n$ sufficiently large,

$$\hat{\boldsymbol{\beta}}_n \overset{d}{\approx} N(\boldsymbol{\beta_0}, \mathcal{I}_n^{-1}(\boldsymbol{\beta_0}))$$

- Conclusion also immediately implies that

$$(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta_0})^\top \mathcal{I}_n(\boldsymbol{\beta_0})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta_0}) \xrightarrow{d} \chi^2_p.$$

- Allows adapting testing/CI developed before using LR and Wald statistics.

In Gaussian linear regression we used sums of squares to measure fit and compare nested models. What about in GLM?

Idea: compare best possible to observed maximised loglikelihood

- Let $\ell_n(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{Y} + \sum_{i=1}^n \gamma(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})$ be the maximised loglikelihood.
- Define the saturated model to be that which has

$$\#\text{parameters} = \#\text{observations}$$

  i.e. where we replace $\boldsymbol{X}_n\boldsymbol{\beta}$ by some unconstrained $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)^\top \in \mathbb{R}^n$. (and thus we also replace $\boldsymbol{x}_i^\top \boldsymbol{\beta}$ by $\eta_i$).
- Let $\hat{\boldsymbol{\eta}}$ be the maximiser of $\ell_n(\boldsymbol{\eta}) = \boldsymbol{\eta}^\top \boldsymbol{Y} + \sum_{i=1}^n \gamma(\eta_i)$ w.r.t. $\boldsymbol{\eta}$.
- Define the saturated loglikelihood as $\ell_n(\hat{\boldsymbol{\eta}})$.

## Definition ((Scaled) Deviance)

$$D = 2\big(\ell_n(\hat{\boldsymbol{\eta}}) - \ell_n(\hat{\boldsymbol{\beta}})\big) = 2\Big((\hat{\boldsymbol{\eta}} - \boldsymbol{X}_n\hat{\boldsymbol{\beta}})^\top \boldsymbol{Y} + \sum_{i=1}^{n}(\gamma(\hat{\eta}_i) - \gamma(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}))\Big)$$

Comments:

- Always $D \geq 0$ (why?)
- Small $D$ implies a good model fit ($\boldsymbol{X}_n\hat{\boldsymbol{\beta}} \approx \hat{\boldsymbol{\eta}}$).
- Large $D$ implies poor fit.
- In Gaussian case: deviance $\equiv$ residual sum of squares (exercise).
- Can now use deviance differences to mimic sum-of-square ratios and construct a GLM variant of the F-test.

- Consider the problem of comparing two nested models:
  - Model $A$: $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ vary freely — MLE $\hat{\boldsymbol{\beta}}^A$
  - Model $B$: for $q < p$, $(\beta_1, \ldots, \beta_q) \in \mathbb{R}^q$ vary freely, but $\beta_{q+1}, \ldots, \beta_p$ are fixed — hence only $q$ free parameters, with MLE $\hat{\boldsymbol{\beta}}^B$

- Model $B$ is *nested within* model $A$: $B$ can be obtained by restrictions on $A$
  - More generally, could have Model $B$ with $\boldsymbol{\beta}$ constrained to vary in a subspace $\mathcal{V}$ of dimension $q < p$, which we can write as $\boldsymbol{\beta} = \boldsymbol{Q}_{p\times q} \underbrace{\boldsymbol{\zeta}}_{q\times 1}$ for $\mathcal{M}(\boldsymbol{Q}) = \mathcal{V}$.

- Likelihood ratio test statistic for comparing the models is

$$2\big(\ell_n(\hat{\boldsymbol{\beta}}^A) - \ell_n(\hat{\boldsymbol{\beta}}^B)\big) = D_B - D_A,$$

and when model $B$ is correct $D_B - D_A \xrightarrow{d} \chi^2_{p-q}$.

Main idea: use deviance instead of sums-of-squares and use final iterate of IWLS to get hat matrix

- Leverage $h_{jj}$ defined as $j$th diagonal element of

$$\boldsymbol{H} = \boldsymbol{V}^{1/2}(\hat{\boldsymbol{\beta}})\boldsymbol{X}_n(\boldsymbol{X}_n^\top \boldsymbol{V}(\hat{\boldsymbol{\beta}})\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^\top \boldsymbol{V}^{1/2}(\hat{\boldsymbol{\beta}}),$$

- Cook statistic now becomes the change in deviance

$$2p^{-1}\left\{\ell(\hat{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}_{-j})\right\},$$

  where $\hat{\boldsymbol{\beta}}_{-j}$ is MLE when $j$th case $(\boldsymbol{x}_j^\top, Y_j)$ is dropped.

- Cook statistic can be approximated by

$$C_j = \frac{h_{jj}}{p(1-h_{jj})} r_{Pj}^2,$$

  where $r_{Pj}$ is standardised Pearson residual (to be defined in next slide).

- Deviance residual:

$$d_j = \operatorname{sgn}(\hat{\eta}_j - \boldsymbol{x}_i^\top \hat{\beta}_j)\Big[2\{\underbrace{\eta_j Y_j + \gamma(\eta_j)}_{\ell_j(\hat{\boldsymbol{\eta}})} - \underbrace{[(\boldsymbol{x}_j^\top \boldsymbol{\beta}) Y_j + \gamma(\boldsymbol{x}_j^\top \boldsymbol{\beta})]}_{\ell_j(\hat{\boldsymbol{\beta}})}\}\Big]^{1/2},$$

for which we note that

$$\sum_{j=1}^{n} d_j^2 = D$$

gives the deviance (in analogy with RSS in Gaussian linear regression).

- Pearson residual:

$$p_j = \frac{Y_i - \gamma'(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})}{\sqrt{\gamma''(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})}} = \frac{Y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \quad \text{so } \boldsymbol{r}_P = \boldsymbol{V}^{-1/2}(\hat{\boldsymbol{\beta}})(\boldsymbol{Y} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}})).$$

- Standardised versions:

$$r_{Dj} = \frac{d_j}{(1 - h_{jj})^{1/2}} \qquad \& \qquad r_{Pj} = \frac{p_j}{(1 - h_{jj})^{1/2}}.$$