Regression Methods

Myrto Limnios

Autumn 2025 - Week 4

Recap on Nested Models/ANOVA

Comparing nested models

Consider the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon.$$

This will always have higher R^2 than the sub-model:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon.$$

- Why? (think of geometry...)
- The question is: is the first model *significantly* better than the second one?
  - ↪ i.e. does the first model explain the variation adequately enough, or should we incorporate extra explanatory variables? Need a quantitative answer.

Gaussian linear model

Model is $y = X\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$. Estimator:

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

Interpretation: $\widehat{y} = X\widehat{\beta} = Hy$ is the projection of y into the column space of X, $\mathcal{M}(X)$. This subspace has dimension p, when X is of full column rank p.
Now for $q < p$ write X in block notation as

$$X = \underset{n\times q}{X_1} \quad \underset{n\times(p-q)}{X_2}.$$

Interpretation: X_1 is built by the first q columns of X and X_2 by the rest.
Similarly write $\beta = (\beta_1 \ \beta_2)^\top$ so that:

$$y = X\beta + \varepsilon = (X_1 \ X_2)\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon.$$

Our question can now be stated as:

- Is $\beta_2 = 0$?

Residual Sums of Squares

Let $H_1 = X_1(X_1^\top X_1)^{-1}X_1^\top$, and $\widehat{y}_1 = H_1 y$, $e_1 = y - \widehat{y}_1$.

Pythagoras tells us that:

$$\underbrace{\|y - \widehat{y}_1\|^2}_{RSS(\widehat{\beta}_1)=\|e_1\|^2} = \underbrace{\|y - \widehat{y}\|^2}_{RSS(\widehat{\beta})=\|e\|^2} + \underbrace{\|\widehat{y} - \widehat{y}_1\|^2}_{RSS(\widehat{\beta}_1)-RSS(\widehat{\beta})=\|e-e_1\|^2}$$

Notice that $RSS(\widehat{\beta}_1) \geq RSS(\widehat{\beta})$ always (think why!)

So the simple idea that we developed week 3 is that : to see if it is worthwhile to include β_2 we will compare how much larger $RSS(\widehat{\beta}_1)$ is to $RSS(\widehat{\beta})$.

- Equivalently, we can look at a ratio like $\{RSS(\widehat{\beta}_1) - RSS(\widehat{\beta})\}/RSS(\widehat{\beta})$
- This is in fact the likelihood ratio test statistic for our hypothesis.
- To construct a test based on this quantity, we need its sampling distributions - cf slides 22-23 week 3.

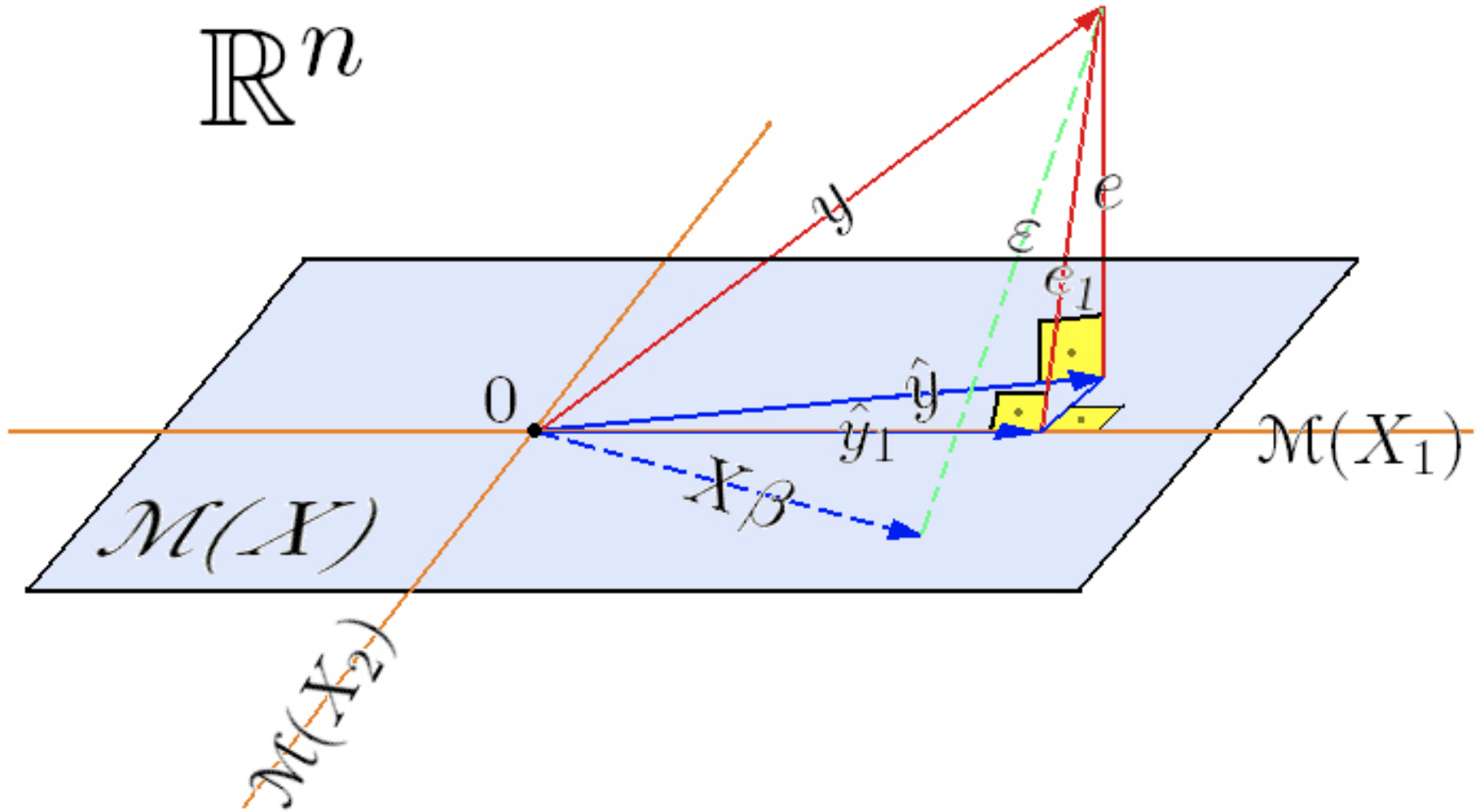


Figure: Geometry Revisited

The Linear Model - Model Building

Goals

- What to do faced with a set of data?
- Two main aims:
  - **understand** (science) — maybe have prior idea/hypotheses on how response depends on explanatory variables. Interpretation is key.
  - **predict/control** (technology) — don't really care how y depends on X. Interpretation not critical (though this describes only prediction in the narrowest of senses).
- There is no reason that a single model will do both, or even that there must be a single 'best' model:
  - maybe two models with different interpretations both fit about equally well, and then future work might aim to choose between them;
  - prediction with a mixture of models might be better than using a single model.

Meta-algorithm

- **Collect** data intended to answer question of interest;
- **examine** data (graphs, look for outliers, problems with sampling scheme);
- **choose/construct** response variable (transformations? independence?);
- **consider** what models are coherent with context of problem (limiting properties, units, similar problems/datasets, covariates that must be included, ...);
- **iterate**:
  - fit models, compare quality of fits;
  - check interpretations of $\widehat{\beta}, \widehat{\sigma}^2$ and
  - check fit (diagnostics, outliers, ...)

  until satisfied; finally
- give **conclusions**—careful interpretation of best model(s) in terms of original problem, consider deficiencies, and explain what extra data might overcome them.

Initial examination of data

- Plot y against covariates, look for outliers, non-constant variance, nonlinearity, etc.
- Plot covariates against each other, look for dependence.
- Try to understand covariates (e.g., dimensions), are transformations needed?
- May need to reduce dimension of X by **regularisation** — many ways to do this (later).

The Linear Model - Variable Selection

Automatic variable selection

- Assume linear model $\mathrm{E}(y) = X\beta$
- 2^p possible subsets of columns of X, plus transformations, ...
- Example: $p = 17$ gives 131072 possible subsets of variables
- Fast algorithms (e.g., branch and bound, `leaps` in `R`) exist visit them all or just subsets (e.g., stepwise), but we need criteria for comparing models.
- Many proposals for model comparison
  1. cross-validation,
  2. information criteria (AIC, $\mathrm{AIC_c}$, BIC, NIC, TIC, ...)
  3. Mallow's C_p,
  4. ...
- Most involve minimising estimated prediction error for future data *like those observed*!

Prediction error

- True model $y \sim (\mu, \sigma^2 I_n)$, we assume (perhaps incorrectly) that $\mu = X\beta$, fit $X_{n\times p}$ and obtain fitted value

$$X\widehat{\beta} = Hy \sim (H\mu, \sigma^2 H).$$

- Terminology
  - the **true model** has $\mu = X\beta$ and all $\beta_r \neq 0$;
  - a **correct model** has $\mu = X\beta$ but some $\beta_r = 0$;
  - a **wrong** model has $\mu \notin \text{span}(X)$;

  so $(I_n - H)\mu = 0$ if the model is true or correct, and $(I_n - H)\mu \neq 0$ if it is wrong.
- The **prediction error** for an independent dataset y_+ with mean vector μ is

$$\Delta = n^{-1}\mathrm{E}\left\{(y_+ - X\widehat{\beta})^{\mathrm{T}}(y_+ - X\widehat{\beta})\right\} = \begin{cases} n^{-1}\mu^{\mathrm{T}}(I-H)\mu + (1+p/n)\sigma^2, & \text{wrong,} \\ (1+q/n)\sigma^2, & \text{true,} \\ (1+p/n)\sigma^2, & \text{correct,} \end{cases}$$

  where $\mathrm{E}(\cdot)$ is over both y_+ and y and $p \geq q = \#\{\beta_r : \beta_r \neq 0\}$ when $\mu \in \text{span}(X)$.
- In principle we should write $\Delta \equiv \Delta(X)$.

Computation of Δ

Let $y \sim (\mu, \sigma^2 I)$ and fit $X\beta$, obtaining fitted value

$$X\widehat{\beta} = Hy \sim (H\mu, \sigma^2 H),$$

where $H\mu = \mu$, i.e., $(I - H)\mu = 0$ if $\mu \in \text{span}(X)$, but otherwise $(I - H)\mu \neq 0$. We have a new data set $y_+ \sim (\mu, \sigma^2 I)$, and we compute the average error in predicting y_+ using $X\widehat{\beta}$, i.e.,

$$\Delta = n^{-1}\mathrm{E}\left\{\|y_+ - \widehat{y}\|^2\right\} =: n^{-1}\mathrm{E}\left\{\|e_+\|^2\right\}.$$

Note that as the trace of a scalar: it is a scalar, and trace is a linear operator,

$$\mathrm{E}(e_+^{\mathrm{T}} e_+) = \mathrm{E}\{\mathrm{tr}(e_+^{\mathrm{T}} e_+)\} = \mathrm{E}\{\mathrm{tr}(e_+ e_+^{\mathrm{T}})\} = \mathrm{tr}\{\mathrm{E}(e_+ e_+^{\mathrm{T}})\} = \mathrm{tr}\{\mathrm{var}(e_+) + \mathrm{E}(e_+)\mathrm{E}(e_+)^{\mathrm{T}}\}.$$

Now as y_+ and y are independent and $\mathrm{var}(X\widehat{\beta}) = \sigma^2 H$, we have

$$y_+ - X\widehat{\beta} \sim (\mu - H\mu, \sigma^2 I + \sigma^2 H),$$

so the computation above gives

$$\mathrm{E}\left\{\|y_+ - X\widehat{\beta}\|^2\right\} = \mathrm{tr}\{\sigma^2(I+H) + (I-H)\mu\mu^{\mathrm{T}}(I-H)\} = \sigma^2(n+p) + \mu^{\mathrm{T}}(I-H)\mu,$$

because $\mathrm{tr}(I + H) = n + p$ and $I - H$ is symmetric and idempotent, giving

$$\Delta = \begin{cases} n^{-1}\mu^{\mathrm{T}}(I-H)\mu + (1 + p/n)\sigma^2, & \text{wrong model,} \\ (1 + q/n)\sigma^2, & \text{true model,} \\ (1 + p/n)\sigma^2, & \text{correct model.} \end{cases}$$

Bias/variance trade-off

- Minimising Δ involves balancing the
  - **bias** $n^{-1}\mu^{\mathrm{T}}(I-H)\mu$, which is reduced by including useful terms in X, and
  - **variance** $(1+p/n)\sigma^2$, which is increased by including useless terms in X.
- We would like to minimise Δ, but it depends on the unknown μ and σ.
- The **cross-validation** estimator of Δ splits the data into X', y' and X^*, y^*, then

Cross-validation with MSE

- for each possible subset $\mathcal{S}$ of columns of X^*:
  - compute $\widehat{\beta}^*_{\mathcal{S}}$ by regressing y^* on $X^*_{\mathcal{S}}$;
  - use $\widehat{\beta}^*_{\mathcal{S}}$ to estimate the prediction error for $\mathcal{S}$ by

$$\widehat{\Delta}_{\mathcal{S}} = (n')^{-1}(y' - X'_{\mathcal{S}}\widehat{\beta}^*_{\mathcal{S}})^{\mathrm{T}}(y' - X'_{\mathcal{S}}\widehat{\beta}^*_{\mathcal{S}});$$

- finally choose the set of columns $\mathcal{S}$ for which $\widehat{\Delta}_{\mathcal{S}}$ is minimised.

This estimator depends on the split, and since $X' \neq X^*$ in general, $\widehat{\Delta}_{\mathcal{S}}$ does not estimate $\Delta_{\mathcal{S}}$, so it would be preferable to use the entire dataset ...

Leave-one-out cross-validation

- Simplest way to use entire dataset is **leave-one-out cross-validation (CV)**, minimising

$$n\widehat{\Delta}_{\mathrm{CV}} = \mathrm{CV} = \sum_{j=1}^{n}(y_j - x_j^{\mathrm{T}}\widehat{\beta}_{-j})^2,$$

  where $\widehat{\beta}_{-j}$ is estimate computed without (x_j, y_j).

- This seems to require n fits, but the lemma below implies that with one fit we have

$$\mathrm{CV} = \sum_{j=1}^{n}\frac{(y_j - x_j^{\mathrm{T}}\widehat{\beta})^2}{(1-h_{jj})^2}.$$

Lemma 9

For a fit $\widehat{y} = Hy$ where H has jth diagonal element h_{jj} and $\widehat{y}_{j,-j}$ is the fitted value for y_j obtained when (x_j, y_j) is dropped,

$$y_j - \widehat{y}_{j,-j} = \frac{y_j - \widehat{y}_j}{1-h_{jj}},$$

and therefore

$$\sum_{j=1}^{n}(y_j - \widehat{y}_{j,-j})^2 = \sum_{j=1}^{n}\frac{(y_j - \widehat{y}_j)^2}{(1-h_{jj})^2}.$$

Proof Lemma 9

- Consider any linear fit $\widehat{y} = Hy$, and note that $\widehat{y}_j = \sum_{i=1}^n h_{ji} y_i$.
- Now suppose we leave out (x_j, y_j) and compute the corresponding (penalized) estimate

$$\widehat{\beta}_{-j} = \text{argmin}_\beta \quad \sum_{i \neq j} (y_i - x_i^{\mathrm{T}} \beta)^2 + \lambda p(\beta),$$

  and fitted value $y_j^* = \widehat{y}_{j,-j} = x_j^{\mathrm{T}} \widehat{\beta}_{-j}$ corresponding to x_j.
- Inserting (x_j, y_j^*) back into the dataset used to compute $\widehat{\beta}_{-j}$ changes nothing, because $(y_j^* - x_j^{\mathrm{T}} \widehat{\beta}_{-j})^2 = 0$ and $p(\beta)$ does not depend on the data. For this new dataset,

$$y_j^* = \sum_{i \neq j} h_{ji} y_i + h_{jj} y_j^* = \sum_{i=1}^n h_{ji} y_i + h_{jj}(y_j^* - y_j) = \widehat{y}_j + h_{jj}(y_j^* - y_j)$$

  so

$$y_j - y_j^* = y_j - \widehat{y}_j + h_{jj}(y_j - y_j^*),$$

  leading to

$$y_j - y_j^* = y_j - \widehat{y}_{j,-j} = \frac{y_j - \widehat{y}_j}{1 - h_{jj}},$$

  and thus to the given formula.

Generalized cross-validation

- Leave-one-out CV can be unstable if some of the h_{jj} are large.
- **Generalised cross-validation (GCV)** replaces all the h_{jj} by their average $\mathrm{tr}(H)/n = p/n$, giving

$$\mathrm{GCV} = \sum_{j=1}^{n} \frac{(y_j - x_j^{\mathrm{T}}\widehat{\beta})^2}{(1-p/n)^2},$$

  and hence

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

  (why? Exercise)

- Often choose the model (i.e. variables) that minimises GCV or CV.
- Note that these only require the second-order assumptions.

Akaike information criterion

1. The above arguments apply only to least squares estimators. More generally, we could aim to minimise the **Kullback–Leibler discrepancy**

$$D(f_\theta, g) = \int \log\left\{\frac{g(y)}{f(y;\theta)}\right\} g(y)\,\mathrm{d}y \geq 0,$$

between **candidate model** $f_\theta \equiv f(y;\theta)$ and true model g, based on $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} g$.

2. Suppose that θ_g minimises $D(f_\theta, g)$ within the family of candidate models, and is estimated by the MLE $\widehat{\theta}$, with log likelihood $\widehat{\ell}$.

3. We suppose there is an independent sample $Y_1^+, \ldots, Y_n^+ \overset{\text{iid}}{\sim} g$ and aim to estimate

$$\mathrm{E}_g\left(\mathrm{E}_g^+\left[\sum_{j=1}^{n} \log\left\{\frac{g(Y_j^+)}{f(Y_j^+;\widehat{\theta})}\right\}\right]\right) = n\mathrm{E}_g\left\{D(f_{\widehat{\theta}}, g)\right\}; \qquad (1)$$

the outer expectation is over the distribution of $\widehat{\theta}$, which is independent of Y^+.

4. After tedious expansions we end up trying to minimise the **Akaike information criterion**

$$\mathrm{AIC} = -2\widehat{\ell} + 2p \qquad (\equiv n \log \mathrm{RSS} + 2p \text{ in linear model}).$$

Reminder on MLE

Because of (2), θ_g depends on g, and differentiating gives the *score equation*

$$0 = \int \frac{\partial}{\partial \theta} \log f(y;\theta_g)\, g(y)\, dy,$$

with $\widehat{\theta}$ determined by the finite-sample version of this,

$$0 = \frac{1}{n} \sum_{j=1}^{n} \frac{\partial}{\partial \theta} \log f(y_j;\widehat{\theta}). \tag{2}$$

Taylor series expansion shows that $\log f(y;\widehat{\theta})$ approximately equals

$$\log f(y;\theta_g) + (\widehat{\theta}-\theta_g)^{\mathrm{T}} \frac{\partial \log f(y;\theta_g)}{\partial \theta} + \frac{1}{2}(\widehat{\theta}-\theta_g)^{\mathrm{T}} \frac{\partial^2 \log f(y;\theta_g)}{\partial \theta \partial \theta^{\mathrm{T}}} (\widehat{\theta}-\theta_g),$$

yielding

$$\widehat{\theta} \doteq \theta_g + \left\{ -\frac{1}{n} \sum_{j=1}^{n} \frac{\partial^2}{\partial \theta \partial \theta^{\top}} \log f(y_j;\theta_g) \right\}^{-1} \left\{ \frac{1}{n} \sum_{j=1}^{n} \frac{\partial}{\partial \theta} \log f(y_j;\theta_g) \right\}.$$

So

$$\widehat{\theta} \;\sim\; N\{\theta_g,\, I_g(\theta_g)^{-1} K(\theta_g) I_g(\theta_g)^{-1}\}.$$

We recognize

$K(\theta_g) = n \int \frac{\partial \log f(y;\theta)}{\partial \theta} \frac{\partial \log f(y;\theta)}{\partial \theta^{\top}}\, g(y)\, dy$, empirical information sandwich variance

$I_g(\theta_g) = -n \int \frac{\partial^2 \log f(y;\theta)}{\partial \theta \partial \theta^{\top}}\, g(y)\, dy$, empirical Fisher information

Derivation of AIC

- Now, consider an independent sample $Y_1^+, \ldots, Y_n^+ \overset{\text{iid}}{\sim} g$ over which we take the expectation over

$$nD(f_{\widehat{\theta}}, g) = n \int \log \left\{ \frac{g(y)}{f(y;\widehat{\theta})} \right\} g(y)\, dy \doteq nD(f_{\theta_g}, g) + \tfrac{1}{2}\mathrm{tr}\left\{ (\widehat{\theta} - \theta_g)(\widehat{\theta} - \theta_g)^{\mathrm{T}} I_g(\theta_g) \right\}$$

  where we have used the fact that the trace of a scalar is itself.

- Using the reminder, the expectation over the distribution of $\widehat{\theta}$ gives its variance matrix, $I_g(\theta_g)^{-1} K(\theta_g) I_g(\theta_g)^{-1}$, and hence

$$n\mathrm{E}_g \left\{ D(f_{\widehat{\theta}}, g) \right\} \doteq nD(f_{\theta_g}, g) + \tfrac{1}{2}\mathrm{tr}\left\{ I_g(\theta_g)^{-1} K(\theta_g) \right\}, \tag{3}$$

  where the second term penalizes the dimension p of θ.

  The first term here is $O(n)$ but the second is $O(p)$.

- If the model is correct, $f_{\theta_g} = g$, $I_g(\theta_g) = K(\theta_g)$ so $\mathrm{tr}\left\{ I_g(\theta_g)^{-1} K(\theta_g) \right\} = p$.

Of course I, K are unknown in practice ...

Derivation of AIC

- To build an estimator, note that $\int \log g(y)\, g(y)\, dy$ is constant and can be ignored. Now $\ell(\widehat{\theta}) = \ell(\theta_g) + \{\ell(\widehat{\theta}) - \ell(\theta_g)\}$, so

$$\begin{aligned} \mathrm{E}_g\left\{-\ell(\widehat{\theta})\right\} &= -\mathrm{E}_g\left\{\ell(\theta_g) + \tfrac{1}{2}W(\theta_g)\right\} \\ &\doteq nD(f_{\theta_g}, g) - \tfrac{1}{2}\mathrm{tr}\left\{I_g(\theta_g)^{-1}K(\theta_g)\right\} - n\int \log g(y)\, g(y)\, dy, \end{aligned}$$

  where we have used the fact that under the wrong model, the Taylor series approximation of the likelihood ratio statistic $W(\theta_g)$ gives

$$2\{\ell(\widehat{\theta}) - \ell(\theta_g)\} \doteq n(\widehat{\theta} - \theta_g)^\top I_g(\theta_g)(\widehat{\theta} - \theta_g),$$

  and the normal distribution of $\widehat{\theta}$ implies that the likelihood ratio statistic has a distribution proportional to χ^2_p, but with mean

$$\mathrm{tr}\{I_g(\theta_g)^{-1}K(\theta_g)\}.$$

- Hence $-\ell(\widehat{\theta})$ tends to underestimate $nD(f_{\theta_g}, g) - n\int \log g(y)\, g(y)\, dy$. On reflection this is obvious, because $\ell(\widehat{\theta}) \geq \ell(\theta_g)$ by definition of $\widehat{\theta}$. As p increases, so will the extent of overestimation.

Derivation of AIC

- An estimator is $-\ell(\widehat{\theta}) + c$, where c estimates $\operatorname{tr}\{I(\theta_g)^{-1}K(\theta_g)\}$. Two possible choices of c are p and $\operatorname{tr}(\widehat{I}^{-1}\widehat{K})$, and these lead to

$$\text{AIC} = 2\{-\ell(\widehat{\theta}) + p\}, \quad \text{NIC} = 2\{-\ell(\widehat{\theta}) + \operatorname{tr}(\widehat{I}^{-1}\widehat{K})\}; \tag{4}$$

- The model is chosen to minimise AIC, say, with the factor 2 putting differences of AIC on the same scale as likelihood ratio statistics. Such criteria are used far beyond random samples, and even in cases where the theory above doesn't work.

- In particular, the maximised log-likelihood for a normal-theory linear model with residual sum of squares RSS can be shown to be

$$-\frac{n}{2}\log(2\pi\widehat{\sigma}) - \frac{n}{2} \equiv -\frac{n}{2}\log \text{RSS} + \text{constants},$$

  which leads to the following formula.

Other model selection criteria

- **'Corrected' AIC** for (normal-theory) regression problems:

$$\text{AIC}_\text{c} \equiv n \log \widehat{\sigma}^2 + n \frac{1 + p/n}{1 - (p+2)/n}.$$

- **Bayes' information criterion**

$$\text{BIC} = -2\widehat{\ell} + p \log n.$$

- **Mallows C_p**:

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

where SS_p is RSS for fitted model and s^2 estimates σ^2.

- When the true model is a candidate and $n \to \infty$,
  - AIC is **inconsistent** — it will not choose the true model with probability one, but tends to pick a more complex model;
  - AIC_c is also inconsistent but gives better results in finite samples;
  - BIC is **consistent** — it chooses the true model with probability $\to 1$.

  These results suppose that the models are fixed, but in practice we also have $p \to \infty$ when $n \to \infty$, because we fit ever more complex models when we have more data.

Simulation experiment

Number of times models were selected using various model selection criteria in 50 repetitions using simulated normal data for each of 20 design matrices. The true model has $p = 3$.

n		Number of covariates						
		1	2	3	4	5	6	7
10	C_p		131	504	91	63	83	128
	BIC		72	373	97	83	109	266
	AIC		52	329	97	91	125	306
	AIC_c	15	398	565	18	4		
20	C_p		4	673	121	88	61	53
	BIC		6	781	104	52	30	27
	AIC		2	577	144	104	76	97
	AIC_c		8	859	94	30	8	1
40	C_p			712	107	73	66	42
	BIC			904	56	20	15	5
	AIC			673	114	90	69	54
	AIC_c			786	105	52	41	16

Stepwise methods

- In principle we might wish to fit all 2^p possible choices of covariates, but in practice this is possible only for 'modest' p, using **leaps** (exhaustive search) or similar methods (or approximations).
- When p is too large for exhaustive searches, we instead consider subsets of the models, using the methods below (or variants).
- **Forward selection**: starting from the model with a constant only,
  1. add each remaining term separately to the current model;
  2. if none of these terms improves the fit, stop; otherwise
  3. update the current model to include the most useful new term; go to 1
- **Backward elimination**: starting from the model with all terms,
  1. if all terms are 'useful', stop; otherwise
  2. update current model by dropping the 'least useful' term; go to 1
- **Stepwise**: starting from an arbitrary model,
  1. consider three options—add a term, delete a term, swap a term in the model for one not in the model;
  2. if model unchanged, stop; otherwise go to 1
- 'Useful' means 'reduces the AIC' (but in the past meant 'is significant using an F test').

Stepwise methods: Comments

- Original formulation of stepwise used F tests (or even arbitrary numbers!) to assess significance, but this finds spurious models.
- Systematic search minimising AIC or similar over all possible models is preferable, but is often infeasible.
- Compare AICs for different models at each step—i.e., use AIC (or AIC_c) as objective function.
- Important not to fixate on a specific model, or assume that there is a single 'best' model, but to consider any models whose AIC is within (say) 2 of the minimum — especially if the interpretations of competing models differ.

The Linear Model - Robustness and Estimating Functions

M-estimation

- The least squares estimates are linear in y and therefore very sensitive to outliers.
- When $y_i \mapsto y_i + c$,

$$\widehat{\beta} = \sum_{j=1}^{n} (X^{\mathrm{T}}X)^{-1} x_j y_j \mapsto \sum_{j=1}^{n} (X^{\mathrm{T}}X)^{-1} x_j y_j + (X^{\mathrm{T}}X)^{-1} x_i c = \widehat{\beta} + (X^{\mathrm{T}}X)^{-1} x_i c,$$

  which could be arbitrarily far from $\widehat{\beta}$.
- Try and fix this by replacing

$$\min_{\beta} \sum_{j=1}^{n} (y_j - x_j^{\mathrm{T}}\beta)^2 \qquad \text{by} \qquad \min_{\beta} \sum_{j=1}^{n} \rho\left\{(y_j - x_j^{\mathrm{T}}\beta)/\sigma\right\},$$

  for function $\rho(\cdot)$ that will give a more robust **M**(aximum likelihood-like)**-estimator**, or equivalently solving the $p \times 1$ system of **estimating equations**

$$\frac{1}{\sigma} \sum_{j=1}^{n} x_j \rho'\left\{(y_j - x_j^{\mathrm{T}}\beta)/\sigma\right\} = X^{\mathrm{T}}\rho' = 0$$

  say, where $\rho'_{n\times 1}$ has jth element $\mathrm{d}\rho(u)/\mathrm{d}u$ for $u = (y_j - x_j^{\mathrm{T}}\beta)/\sigma$.

Choice of ρ

- Choose $\rho(u)$ to have desirable properties, e.g., to downweight outliers:

$$\begin{aligned}
\rho(u) &= u^2/2 \quad \text{(normal errors)},\\
\rho(u) &= |u| \quad \text{(Laplace errors)},\\
\rho(u) &= \nu \log(1 + u^2/\nu)/2 \quad (t_\nu \text{ errors}),\\
\rho(u) &= \begin{cases} u^2/2, & |u| < c,\\ c(2|u| - c)/2, & \text{otherwise,} \end{cases} \quad \text{(Huber function).}
\end{aligned}$$

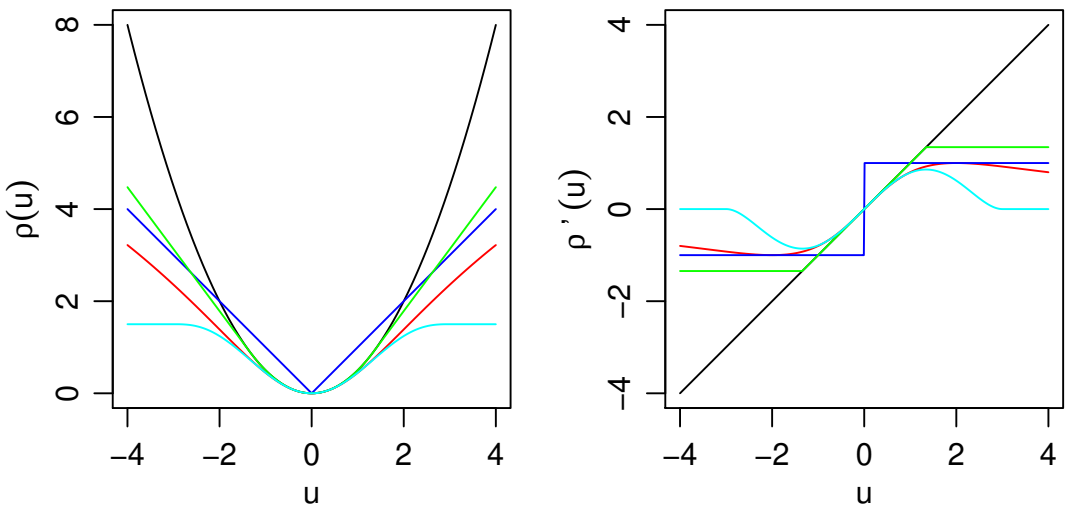
- The function $\rho'(u)$ is also called the **influence function** of the estimator, as its value determines what influence an observation at u has on the estimator:
  - Huber $\rho'(u)$ is bounded,
  - t_ν function is bounded and **redescending**, as $\lim_{u\to\pm\infty} \rho'(u) = 0$;
  - Tukey's **biweight**

$$\rho'(u) = u\left\{1 - (u/c)^2\right\}^2 I(|u| < c),$$

which gives $\rho'(u) = 0$ when $|u| > c$, is also redescending, giving no weight to observations outside $\pm c$.

ρ and ρ'

Functions ρ and ρ' for least squares (black), t_5 (red), Laplace (blue), Huber (green) and biweight (cyan) estimators.



Estimation

- We need to solve

$$X^{\mathrm{T}}\rho' = 0,$$

where ρ' has jth element

$$\sigma^{-1}\rho'\{(y_j - x_j^{\mathrm{T}}\beta)/\sigma\} \propto \frac{\rho'\{(y_j - x_j^{\mathrm{T}}\beta)/\sigma\}}{y_j - x_j^{\mathrm{T}}\beta} \times (y_j - x_j^{\mathrm{T}}\beta) = w_j(\beta,\sigma)(y_j - x_j^{\mathrm{T}}\beta),$$

say, so we write the estimating equation as

$$X^{\mathrm{T}}W(y - X\beta) = 0,$$

with $W = \mathrm{diag}\{w_1(\beta,\sigma), \ldots, w_n(\beta,\sigma)\}$.

- We use **iterative weighted least squares**: choose some initial $\tilde{\beta}$ and σ, then iterate to convergence the steps
  - compute W using the current $\tilde{\beta}$,
  - compute the weighted least squares estimate,

$$\tilde{\beta} = (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}Wy.$$

- Estimate σ using median absolute deviation of residuals $y_j - x_j^{\mathrm{T}}\tilde{\beta}$ at each iteration, or similar robust scale estimate.

M-estimator variance

- Estimator $\tilde{\beta}$ is solution to $p \times 1$ system of equations

$$g(y; \beta) = X^{\mathrm{T}} \rho' = 0.$$

- Can show that if the estimating function g is **unbiased**, i.e.

$$\mathrm{E}\left\{g(Y; \beta)\right\} = 0, \qquad \text{for any } \beta,$$

then under mild regularity conditions

$$\tilde{\beta} \dot{\sim} \mathcal{N}_p \left(\beta, \mathrm{E}\left\{ -\frac{\partial g(Y;\beta)}{\partial \beta^{\mathrm{T}}} \right\}^{-1} \mathrm{var}\left\{g(Y;\beta)\right\} \mathrm{E}\left\{ -\frac{\partial g(Y;\beta)}{\partial \beta^{\mathrm{T}}} \right\}^{-1} \right).$$

This is another **sandwich** variance matrix, with

$$\mathrm{E}\left\{ -\frac{\partial g(Y;\beta)}{\partial \beta^{\mathrm{T}}} \right\} = X^{\mathrm{T}} W_1 X, \qquad \mathrm{var}\left\{g(Y;\beta)\right\} = X^{\mathrm{T}} W_2 X,$$

so if $W_1 = A(\sigma) I_n$, $W_2 = \sigma^2 B(\sigma) I_n$, then

$$\mathrm{var}(\tilde{\beta}) \doteq \sigma^2 (X^{\mathrm{T}} X)^{-1} \times B(\sigma)/A(\sigma)^2.$$

Proof of sandwich matrix

- The $p \times 1$ estimating function is

$$g(y;\beta) = \sum_{j=1}^{n} x_j \rho'\left(\frac{y_j - x_j^{\mathrm{T}}\beta}{\sigma}\right),$$

and unbiasedness implies that if the individual densities are $\sigma^{-1} f\{(y_j - x_j^{\mathrm{T}}\beta)/\sigma\}$, then

$$0 = \mathrm{E}\left\{g(y;\beta)\right\} = \sum_{j=1}^{n} x_j \int \rho'\left(\frac{y_j - x_j^{\mathrm{T}}\beta}{\sigma}\right)\sigma^{-1} f\left(\frac{y_j - x_j^{\mathrm{T}}\beta}{\sigma}\right) \mathrm{d}y_j = X^{\mathrm{T}} a_{n\times 1},$$

say, where a_j is the jth integral above, and setting $u = (y_j - x_j^{\mathrm{T}}\beta)/\sigma$ shows that all the a_j equal

$$\int \rho'(u)\, f(u)\, \mathrm{d}u = 0; \tag{5}$$

this is true by symmetry if the error distribution and ρ' are symmetric around the origin. Now

$$\frac{\partial g(y;\beta)}{\partial \beta^{\mathrm{T}}} = -\frac{1}{\sigma}\sum_{j=1}^{n} x_j x_j^{\mathrm{T}} \rho''\left(\frac{y_j - x_j^{\mathrm{T}}\beta}{\sigma}\right),$$

whose expectation is (using the same transformation)

$$\begin{aligned}
\mathrm{E}\left\{\frac{\partial g(y;\beta)}{\partial \beta^{\mathrm{T}}}\right\} &= -\frac{1}{\sigma}\sum_{j=1}^{n} x_j x_j^{\mathrm{T}} \mathrm{E}\left\{\rho''\left(\frac{Y_j - x_j^{\mathrm{T}}\beta}{\sigma}\right)\right\} \\
&= -\frac{1}{\sigma}\sum_{j=1}^{n} x_j x_j^{\mathrm{T}} \int \rho''(u) f(u)\, \mathrm{d}u = -\frac{1}{\sigma} X^{\mathrm{T}} X A(\sigma), \quad \text{say.}
\end{aligned}$$

- The components of these sums are independent, so

$$\operatorname{var}\{g(Y;\beta)\} = \operatorname{var}\left\{\sum_{j=1}^{n} x_j \rho'\left(\frac{Y_j - x_j^{\mathrm{T}}\beta}{\sigma}\right)\right\} = \sum_{j=1}^{n} x_j x_j^{\mathrm{T}} \operatorname{var}\left\{\rho'\left(\frac{Y_j - x_j^{\mathrm{T}}\beta}{\sigma}\right)\right\},$$

where the substitution $u = (y_j - x_j^{\mathrm{T}}\beta)/\sigma$ and (5) show that the variance term can be written as

$$\operatorname{var}\left\{\rho'\left(\frac{Y_j - x_j^{\mathrm{T}}\beta}{\sigma}\right)\right\} = \int \rho'(u)^2 f(u)\,\mathrm{d}u = B(\sigma).$$

- The sandwich variance formula is therefore

$$\left\{-\frac{1}{\sigma}X^{\mathrm{T}}XA(\sigma)\right\}^{-1} X^{\mathrm{T}}XB(\sigma)\left\{-\frac{1}{\sigma}X^{\mathrm{T}}XA(\sigma)\right\}^{-1} = (X^{\mathrm{T}}X)^{-1} \times \frac{\sigma^2 B(\sigma)}{A(\sigma)^2}.$$

The variance of the LSE is $\operatorname{var}(Y_j)(X^{\mathrm{T}}X)^{-1}$, so the asymptotic relative efficiency of the M-estimator based on ρ and the LSE is

$$\frac{\operatorname{var}(Y_j)}{\sigma^2} \times \frac{A(\sigma)^2}{B(\sigma)}.$$

- As a check on this, note that for the normal distribution $\rho'(u) = u$, $f(u) = (2\pi)^{-1}e^{-u^2/2}$, so $A(\sigma) = B(\sigma) = 1$, which gives ARE of 1.
  If we take $\rho'(u) = \text{sign}(u)$ with the normal density, we have $B(\sigma) = 1$, $A(\sigma) = -2/(2\pi)^{1/2}$, so the sandwich variance formula gives $\sigma^2(X^{\mathrm{T}}X)^{-1}\pi/2$.

  So using the ρ-function corresponding to the Laplace distribution when the data are in fact normally distributed leads to an estimator which is $\pi/2 \approx 1.57$ times more variable than would be the case if the appropriate ρ-function were used.

- If we take the ρ-function $\rho'(u) = u$ corresponding to the normal density, and the errors are in fact Laplace, $g(u) = (1/2)e^{-|u|}$, we have

$$A(\sigma) = \int (-1) f(u)\,\mathrm{d}u = 1, \quad B(\sigma) = \int u^2 f(u)\,\mathrm{d}u = 2$$

  and the asymptotic relative efficiency is $1/2$.

Efficiency

- Efficiency of M-estimators of β relative to LSEs of β is

$$\frac{\mathrm{var}(Y_j)}{\sigma^2} \times \frac{A(\sigma)^2}{B(\sigma)};$$

  for example, the Huber estimator is 95% efficient if $c = 1.345$.
- In practice need to balance robustness and efficiency, increasing the latter by increasing c.
- High numbers of outliers can wreck M-estimators.
- Highly robust **least trimmed squares** estimators obtained by minimising

$$\sum_{j=1}^{q} (y_j - x_j^{\mathrm{T}}\beta)^2_{(j)},$$

  where $q = \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$.

Quantile regression

- The Laplace distribution has

$$\rho(u) = |u| = uI(u \geq 0) - uI(u < 0),$$

and for continuous Y, the solution to $\mathrm{E}\{\rho'(Y - \theta)\} = 0$ is the median of Y. Hence

$$\operatorname{argmin} \sum_{j=1}^{n} \rho(y_j - x_j^{\mathrm{T}}\beta)$$

estimates the median of y as a linear function of $X\beta$.

- **Quantile regression** takes $\tau \in (0, 1)$ and uses the **check function**

$$\rho_\tau(u) = \tau uI(u \geq 0) - (1 - \tau)uI(u < 0);$$

then

$$\tilde{\beta}_\tau = \operatorname{argmin} \sum_{j=1}^{n} \rho_\tau(y_j - x_j^{\mathrm{T}}\beta)$$

estimates the τ quantile of y as a linear function of $X\beta$.

- For numerical purposes it may be better to round the boundaries of ρ.
- Note that $\rho''_\tau(u) = 0$, so it's better to bootstrap to find $\mathrm{var}(\tilde{\beta}_\tau)$.

Expectile regression

- Quantile regression can be used to estimate value-at-risk in finance settings, but it has the drawback of just counting how many residuals are above/below the quantile.

- **Expectile regression** extends the LSE in the same way, taking

$$\rho_\tau(y-\theta) = \eta_\tau(y-\theta) - \eta_\tau(y), \quad \eta_\tau(u) = |I(u \leq 0) - \tau| u^2,$$

so $\tau = 1/2$ gives the LSE, while taking $\tau > 1/2$ leads to a more general form of LSE, with good properties for risk estimation in finance applications (coherent elicitable risk measure).