♠ Multiple linear regression models

- Extend straight line regression model to use more than one predictor gene
  (dependent variable) response gene y (e.g., ER)
  (independent variable, explanatory variable) predictor genes $x_1, x_2, ..., x_p$
- Measurement error model: repeat values $i = 1, \ldots, n$,
  - independent expression levels on n tumors

$$y_i = \alpha + \sum_{r=1}^{p} \beta_r x_{r,i} + \epsilon_i$$

  $x_{r,i}$ is expression of gene r on array i
- Model "explains" variability in response y "due to" p genes
- Non-causal, purely empirical
- Predictive validity: fit model and test in new cases

- Interpretation: β_r measures change in expected response with a unit change in predictor x_r
- Value and interpretation of β_r depends, sometimes critically, on which other genes/predictors are in specified model

- Analysis and inference:
  - Estimate parameters $(\alpha, \beta_1, ..., \beta_p, \sigma^2)$
  - Predict new ("future") responses ...

♠ Notation: Matrices and vectors

- Intercept term $\alpha = \beta_0 = \beta_0 x_0$ with $x_0 = 1$ ("dummy" gene with constant expression)
- Revise earlier notation for $\mathbf{x}_i, \mathbf{X}$ to include dummy/intercept
- $(p+1) \times 1$ column vector

$$\mathbf{x}_i = \begin{pmatrix} 1 \\ x_{1,i} \\ x_{2,i} \\ \vdots \\ x_{p,i} \end{pmatrix}$$

- Regression parameter vector

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

- Model is $y_i = \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_i$
- Expression data in $(p+1) \times n$ matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$
- Response variable and errors in $n \times 1$ vectors

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

- Model in matrix form:

$$\mathbf{y} = \mathbf{X}'\beta + \boldsymbol{\epsilon}$$

♠ Least squares fitting

- For any chosen $\boldsymbol{\beta}$,

$$Q(\beta) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\beta)^2$$

  measures "fit" of chosen line $\mathbf{x}_i\boldsymbol{\beta}$ to response data
- Choose $\hat{\boldsymbol{\beta}}$ to *minimise* $Q(\boldsymbol{\beta})$
- Least squares estimates (LSE)
- Fitted least squares line: $\hat{y} = \mathbf{X}'\hat{\boldsymbol{\beta}}$
- Residuals: $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ with elements $e_i = y_i - \hat{y}_i$
  - what is left 'unexplained' in response data
  - estimates of ϵ_i

♠ LSE formulæ:

- $$\hat{\beta} = \mathbf{V}\mathbf{X}\mathbf{y} \quad \text{and} \quad \mathbf{V} = (\mathbf{X}\mathbf{X}')^{-1}$$

  (note: standard notation uses $\mathbf{X}'$ in place of $\mathbf{X}$ in many statistics books)
- Elements of $\hat{\boldsymbol{\beta}}$ measure relationships between the predictors and responses, in the context of all other predictor variables used in the model
- Values depend on the other predictors in the model, and differ in different models, paralleling changing interpretation of β_i parameters
- Center variables so that they have zero-mean, by subtracting sample mean (for each gene) before modelling. Good for numerical stability. One implication is that $\hat{\beta}_0 = 0$

♠ Collinearity of predictors

- Imagine a predictor x_r that is highly positively correlated with response y, so has a high positive regression coefficient estimate in the linear model using only that predictor. Fitting it in a more elaborate model with other xs changes things, often in unpredictable ways. The estimate $\hat{\boldsymbol{\beta}}_r$ may be negligible, even negative. This will be experienced when the multiple predictors are correlated, and is due to other predictors dominating in explaining the response. This is called *collinearity* of predictors, and is the norm rather than the exception. At the other extreme predictors are *orthogonal* if $\mathbf{X}\mathbf{X}'$ is diagonal (so that they are uncorrelated), and in this case coefficient estimates do not depend on which other predictors are used in the model.

♠ Uncertainty and Significance of Predictors

- Standard inference results give standard errors (SDs) for each coefficient, say v_j for $\hat{\beta}_j$, such that that symmetric error bars have the form of $\hat{\beta}_j \pm c.v_j$ for appropriate constants c. This is a probability interval estimate of β_j ("confidence interval"). The regression function in matlab provided in this class plots 95% intervals.
- Similarly, standard inference gives 'probability' levels (or significance levels) for each coefficient. Predictors with non-significant parameters might be dropped from the model (see below).

♠ Simple stepwise fitting

- Choose a set of most highly correlated predictor variable and fit all of them.
- Look at the probability levels for each coefficient; if all are small (i.e., significant), stop, otherwise remove the predictor that is least significant, and repeat.
- Simple 'backward selection' procedure – an old, fairly crude method with some short-comings, but a standard method and a simple start on the difficult and pressing problem of selecting useful and relevant predictor variables in multiple regression models.

♠ Issues, concerns

- Many selection procedures: good properties, bad ones
  (e.g., backward/forward selection, AIC, BIC, etc)
- Purely data based: genes may be biologically relevant, in terms of networks, but 'insignificant' due to idiosyncacies of the data set
- Small sample issues: p should be 'small' relative to sample size n
- Over-fitting concerns: too many predictors, too few samples
- Wide interval estimates, wide prediction intervals in cases of small samples
- Bayesian methods of stochastic regularisation can improve predictions
  (see binary regression models later, where Bayesian methods are used)

♠ Networks

- Regression models help identify which genes are useful predictors of others, in terms of expression levels
- Repeat with various genes selected as response variables
- One way of thinking – empirically – about network relationships
  (more refined methods would use Bayesian statistical methods - Bayes' networks)

♠ Prediction

- Future response value y_{n+1} to be predicted at future predictors $\mathbf{x}_{n+1}$
- New tumor sample, etc
- Predictive validity of regression model: How does the model stand up in out-of-sample prediction?
- Standard inference theory gives predictions in terms of
  - fitted/estimated/predicted value: $\hat{y}_{n+1} = \mathbf{x}'_{n+1}\hat{\boldsymbol{\beta}}$ and corresponding standard error