Special Matrix	Symmetric matrix A is +ve/(-ve) definite if for any vector x, $x^TAx > (<)0$. ≥ (non -ve); ≤ (non +ve) definite If A is a sq matrix and $A^TA = I$, then A is orthogonal matrix -> rows/cols of A form orthonormal basis for R^n -> all rows/cols are pairwise orthogonal (u · v = 0) & all rows/cols are unit vector
	If AB and BA are both compatible -> tr(AB) = tr(BA) & AB and BA have the same non-zero eigenvalues If A is non-negative definite, $\max_x \frac{x^TAx}{x^Tx} = \lambda_{max}(A)$, where λ_{max} denotes largest eigenvalue If A is non-negative definite and B is non-singular (i.e. has inverse, det(B) = 0), $\max_x \frac{x^TAx}{x^TBx} = \lambda_{max}(AB^{-1})$
Descriptive quantities	Var of X is $\sigma^2 = E(X - E(X))^2 = \int (x - E(X))^2 dF(x)$. Sample var is $\frac{1}{n}\sum_{i=1}^n (x_i - \bar{X})^2$ Let Y be another r.v. w observed sample $y_1,..., y_n$. Covariance btw X and Y is Cov(X, Y) = E[(X - E(X))(Y - E(Y))] If Cov(X, Y) = 0, then X and Y are un-correlated. Sample covariance is $\frac{1}{n}\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})$ Suppose X has a cts dist. The (1 - α)-quantile (upper α-quantile), q_α is defined s.t. F(q_α) = 1 - α or 1 - F(q_α) = α
	Mean: $E\left(\sum_{j=1}^m c_jX_j\right) = \sum_{j=1}^m c_jE(X_j)$. Var: Var(X) = E($X^2$) - $[E(X)]^2$ $Var\left(\sum_{j=1}^m a_jX_j\right) = \sum_{i=1}^m \sum_{j=1}^m a_ia_jCov(X_i, X_j)$ in general; $Var\left(\sum_{j=1}^m a_jX_j\right) = \sum_{j=1}^m a_j^2Var(X_j)$ if X_j's are un-correlated Covariance: Cov(X, Y) = E(XY) - E(X)E(Y). $Cov\left(\sum_{i=1}^m a_iX_i, \sum_{j=1}^n b_jY_j\right) = \sum_{i=1}^m \sum_{j=1}^m a_ib_jCov(X_i, Y_j)$
	Sample var: $\frac{1}{n}\sum_{i=1}^n (x_i)^2$ - $(\bar{X})^2$. Sample cov: $\frac{1}{n}\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y}) = \frac{1}{n}\sum_{i=1}^n (x_iy_i)$ - $\bar{X}\bar{Y}$
	Let A, B be constant matrics and **b, c** be vectors. Let X, Y be random vectors. E(AX + **b**) = AE(X) + **b** where E(X) = $(E(X_1), ..., E(X_p))^T$ Var(AX + **b**) = AVar(X)A^T where Var(X) = Cov(X_i, X_j) = E(X - E(X))(X - E(X))^T. Note Cov(X_i, X_i) = Var(X_i) Cov(AX + **b**, BX + **c**) = AVar(X)B^T. Cov(AX + **b**, BY + **c**) = ACov(X, Y)B^T where Cov(X, Y) = Cov(X_i, Y_j) = E(X - E(X))(Y - E(Y))^T
Matrix op on Exp and Var	r.v. X has uni-variate normal dist if pdf is $f(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$ where μ is the mean and σ^2 is the var of the dist Normal dist is denoted by N(μ, σ^2). If μ = 0, σ^2 = 1, it is the standard normal dist. Normal dist is symmetric about its mean If X ~ N(μ, σ^2), then Z = (X - μ)/σ ~ N(0,1) := standardisation
Uni-variate Normal dist	Let **Y** = $(Y_1, ..., Y_m)^T$ be random vector whose components are iid standard normal variables. Let A be q x m constant matrix and $\boldsymbol{\mu}$ a constant vector. Dist of **X** = A**Y** + $\boldsymbol{\mu}$ is a q-dimensional multivariable normal dist w mean E(**X**) = $\boldsymbol{\mu}$ and var matrix ∑ = AA^T, and denoted by N($\boldsymbol{\mu}$, ∑) $\Sigma = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_q) \\ Cov(X_2, X_1) & Var(X_2) & \cdots & Cov(X_2, X_q) \\ \cdots & \cdots & \cdots & \cdots \\ Cov(X_q, X_1) & Cov(X_q, X_2) & \cdots & Var(X_q) \end{pmatrix}$. Normal dist is uniquely determined by its mean and var matrix ∑ pdf of N($\boldsymbol{\mu}$, ∑) is $f(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi\lvert\Sigma\rvert)^{q/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$ If X is a multivariate normal vector, then for any constant matrix B, B**X** has a multivariate normal dist N(B$\boldsymbol{\mu}$, B∑B^T) ∀ constant vector **c**, the LC $\mathbf{c}^T\mathbf{X}$ has a univariate normal dist N($\mathbf{c}^T\boldsymbol{\mu}$, $\mathbf{c}^T\Sigma\mathbf{c}$) and any component of **X** is an univariate normal var If **X**~N($\boldsymbol{\mu}$, ∑), then $\mathbf{Z} = \Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$~N(**0**, I), i.e. components of **Z** are iid N(0, 1) variables
Multi-variate normal dist	Let **Z** = $(Z_1,..., Z_m)^T$. Suppose Z_j's are iid N(0,1) variables. Dist of $\mathbf{Z}^T\mathbf{Z} = \sum_{j=1}^m Z_j^2$ is called the χ^2-dist w d.f. m and denoted by χ_m^2 Suppose Z~N(0,1), U~χ_m^2, Z and U are indep. The dist of Z/($\sqrt{U/m}$) is called the t-dist w d.f. m and denoted by t_m Suppose U~χ_m^2, V~χ_n^2, U and V are indep. Dist of $\frac{U/m}{V/n}$ is called the F-dist w d.f. m and n and denoted by $F_{m,n}$ If **X**~N($\boldsymbol{\mu}$, ∑), then W = $(\mathbf{X} - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$~ χ_m^2 If Z~N(0,I), A is symmetric and idempotent, then $\mathbf{Z}^TA\mathbf{Z}$~χ_r^2, where r = rk(A) = tr(A)
χ^2, t and F - dist	Let θ be param of interest. Suppose θ = g(m_1, m_2,...) (1st moment, 2nd moment,...). Then MME of θ is obtained by replacing the theoretical moments in the fn w the corresponding sample moments, i.e. $\hat{\theta}_{MME}$ = g($\hat{m}_1$, $\hat{m}_2$,...) For any dist, the var $\sigma^2 = m_2 - m_1^2$ (E(X^2) - $[E(X)]^2$), its MME is given by $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{X}^2$. MME is not unique.
Mtd of Moment estimation (MME)	If X has pdf f(x, θ), given the observation $x_1,...,x_n$ of a random sample, the log likelihood fn of θ is defined as $\ell(\theta) = \sum_{i=1}^n ln\ f(x_i, \theta)$. The MLE of θ is value of θ that maximizes the log-likehood fn. E.g. let ($x_1,...,x_n$) be observation of random sample from N(μ, σ^2). Likelihood fn: $\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2}$ (μ is prediction here, i.e. $x_i - \mu$ = y_i - $\hat{y}_i$) The log likelihood fn of (μ, σ^2) is then $\ell(\mu, \sigma^2) = \sum_{i=1}^n log\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2$ MLE of μ and σ^2 are obtained by maximizing $\ell(\mu, \sigma^2)$ and are given by $\hat{\mu} = \bar{X}$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{X})^2$
Maximum likelihood estimation (MLE)	Null hypothesis H_0 and alternative hypothesis H_1. The 2 hypotheses are mutually exclusive A test statistic, T(x) is used. If T(x) ≥ c for a predetermined constant c -> reject H_0. If not -> don't reject H_0 To control the type I error rate (reject H_0 when H_0 true) at a given level α, i.e. choose c s.t. $P(T(\mathbf{X}) \geq c \mid H_0) \leq \alpha$ -> reject H_0 H_0 should be s.t. type I error is more serious. If still not clear, H_0 should be a well-established theory OR opp of new guess

Simple Linear Regres-sion	Correlation ≠ causation. Correlation coefficient can only be btw -1 and 1, i.e. -1 ≤ ρ_{xy} ≤ 1 Pearson's correlation, the theoretical correlation is $\rho_{xy} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$, the sample correlation is $\hat{\rho}_{xy} = \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2 \sum_{i=1}^n (y_i-\bar{y})^2}}$ = corr(X, Y) If ρ_{xy} > 0: move in same dir. If ρ_{xy} < 0: move in opp dir. Both DO NOT imply a causal r/s If ρ_{xy} = 0, X and Y have no LINEAR r/s. X and Y could have other kinds of r/s (e.g. quadratic)
	Y is the *response variable*. X = *covariate*/predictor A simple regression model (SRM) is Y = $\beta_0 + \beta_1$X + ϵ, E(Y) = $\beta_0 + \beta_1$X is called the regression function ($E(Y \mid X)$ more technically correct way to write it) W observations (x_i, y_i), observed simple LRM is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, i = 1,...,n. Error term is diff for ea observation
	Assumptions of simple linear regression model (LRM): 1. x_i and ϵ_i are indep. Note indep → un-correlated, ↚ not necessary. (Unless normal var, then indep <-> un-correlated) 2. ϵ_i have mean 0 3. ϵ_i are pairwise un-correlated, i.e. Cov(ϵ_i, ϵ_j) = 0 4. ϵ_i have common variance σ^2 (Homogeneity) 5. ϵ_i have a Normal dist (Normality) LRM is then called a Normal LRM. Assumptions can be simply stated, ϵ_i, iid ~N(0, σ^2) in addition to (1)
	From Y = $\beta_0 + \beta_1$X + ϵ, 1) E(Y) = $\beta_0 + \beta_1$E(X) 2) Cov(X, Y) = β_1Var(X)

So $\beta_1 = \frac{Cov(X,Y)}{Var(X)} = \rho_{xy}\frac{\sigma_y}{\sigma_x}$ and $\beta_0 = \mu_y - \beta_1\mu_x$ (exact values of β_0, β_1)

(β_0, β_1) will then be the soln to $\min\limits_{b_0,b_1} E(Y - b_0 - b_1X)^2$ (i.e. error term minimised)

So regression fn $\beta_0 + \beta_1$X is best linear approximation of X to Y

Estimation of LRM

LSE of Property (β_0, β_1) minimizes $E(Y - b_0 - b_1X)^2$ gives rise to the least square estimation (LSE), which estimate the params by minimizing the sum of squares, Q = $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1x_i)^2$ = SSE

To find LSE of β_0, β_1, $\frac{\partial Q}{\partial\beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1x_i)$, and $\frac{\partial Q}{\partial\beta_1} = -2\sum_{i=1}^{n}x_i(y_i - \beta_0 - \beta_1x_i)$ and equate both to 0

Summary: $\hat\beta_0 = \bar{Y} - \beta_1\bar{X}$, $\hat\beta_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{X})(y_i-\bar{Y})}{\sum_{i=1}^{n}(x_i-\bar{X})^2}$, where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n}x_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}y_i$

LSE of σ^2 = E(ϵ^2). $\hat{y}_i = \hat\beta_0 + \hat\beta_1x_i$

SSE (sum of square of error) = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}{e_i}^2$ and σ^2 estimated by $\hat\sigma^2$ = s² = SSE/(n-2)

Estimator of σ is $\hat\sigma$ is called the residual standard error

Fitted/estimated regression fn: $\hat{Y} = \hat\beta_0 + \hat\beta_1$X

Fitted values: $\hat{y}_i = \hat\beta_0 + \hat\beta_1x_i$, i = 1,...,n

Residuals: $e_i = y_i - \hat{y}_i$, i = 1,...,n

Value $\hat{Y}^* = \hat\beta_0 + \hat\beta_1$X* is called the predicted value of a response at X*

Variation of Y is estimated by total sum of squares: SST = SSR + SSE

SSE is variation explained by X; regression sum of square

SSE is variation caused by random errors; residual sum of squares. Note $y_i = \hat{y}_i + \epsilon_i$

ANOVA table for SRM

source	df (deg of freedom)	SS (sum of sq)	MS (mean of sq)	F (f ratio)
Regression	1	SSR = $\sum_{i=1}^{n}(\hat{y}_i - \bar{Y})^2$	MSR = SSR/1	MSR/MSE
Error	n-2	SSE = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	MSE = SSE/(n-2)	
Total	n-1	SST = $\sum_{i=1}^{n}(y_i - \bar{Y})^2$		

Coefficient of determination, R² = proportion of variation in Y explained by X. Measures strength of correlation btw Y and X

Adjusted R²: Less biased estimate is given by $R_a^2 = \frac{n-1}{n-p-1}$R² - $\frac{p}{n-p-1}$ where p is num of predictors and equals 1 for simple LRM

R² is strictly increasing as p increases. But R_a^2 does not necessarily incr as p incr

$\hat\beta_0 = \bar{Y} - \beta_1\bar{X}$, $\hat\beta_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{X})(y_i-\bar{Y})}{\sum_{i=1}^{n}(x_i-\bar{X})^2}$, SSR = $\beta_1^2\sum_{i=1}^{n}(x_i - \bar{X})^2$, R² = corr (Y, $\hat{Y}$)² = corr(Y, X)² = $\rho_{xy}^2 = \frac{\beta_1^2\sigma_x^2}{\sigma_y^2} = \frac{\beta_1^2\sigma_x^2}{\beta_1^2\sigma_x^2 + \sigma^2} = \frac{SSR}{SST}$

Summary results explained:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	$\hat\beta_0$	s($\hat\beta_0$)	$\frac{\hat\beta_0}{s(\hat\beta_0)} = T_{\beta_0}$	p-value for 2-sided test on $\hat\beta_1$
X	$\hat\beta_1$	s($\hat\beta_1$)	$\frac{\hat\beta_1}{s(\hat\beta_1)} = T_{\beta_1}$	p-value for 2-sided test on $\hat\beta_1$
Residual standard error:			$\hat\sigma = \sqrt{MSE}$	
Multiple R-squared:		R² = $\frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{MSR}{SST}$	Adjusted R-squared:	R_a^2
F-statistic:		MSR/MSE = $(T_{\beta_1})^2$	p-value	p-value of the significant F test

$\hat{Y} = \hat\beta_0 + \hat\beta_1$X. Using R², can explain how much of variation of Y is due to X

$\hat\beta_1$ is estimated amt of change in expectation of Y when X incr by an unit amt

Can predict Y* w new observation X* using $\hat\beta_0 + \hat\beta_1$X*

Fitted regression function: $\hat{Y} = \hat\beta_0 + \hat\beta_1$X. We can say Y incr/decr as X incr. When X incr by 1 unit, Y incr/decr by $\hat\beta_1$

X explains about (R²*100)% of the variation in Y

Theoretical Properties of LSE

Unbiasedness of s² / $\hat\sigma^2$ as an estimator of σ^2: Thus Es² = σ^2

Properties of estimated/fitted regression fn:

The estimated regression fn $\hat{Y} = \hat\beta_0 + \hat\beta_1$X is an unbiased estimator of EY = $\beta_0 + \beta_1$X, i.e. E($\hat{Y}$) = E(Y)

var($\hat{Y}$) = $\left[\frac{1}{n} + \frac{(X-\bar{X})^2}{\sum_{i=1}^{n}(x_i-\bar{X})^2}\right]\sigma^2$. OR using $\hat{Y} = \hat\beta_0 + \hat\beta_1$X, var($\hat{Y}$) = var($\hat\beta_0$) + 2XCov($\hat\beta_0$, $\hat\beta_1$) + X²var($\hat\beta_1$)

For the prediction $\hat{Y}$ at a new value X, the prediction mean square error is: E(Y-$\hat{Y}$)² = var(ϵ) + var($\hat{Y}$) = $\left[1 + \frac{1}{n} + \frac{(X-\bar{X})^2}{\sum_{i=1}^{n}(x_i-\bar{X})^2}\right]\sigma^2$

Dist of LS estimators: Normality condition is assumed. $\hat\beta_0$, $\hat\beta_1$ are LC of Y_1, ..., Y_n

By property of normal dist, $\hat\beta_0$, $\hat\beta_1$ are normally distributed: $\hat\beta_0$~N$\left(\beta_0, \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(x_i-\bar{X})^2}\right]\sigma^2\right)$, $\hat\beta_1$~N$\left(\beta_1, \left[\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{X})^2}\right]\right)$

(n-2)$\frac{s^2}{\sigma^2}$ (where s² = $\hat\sigma^2$) ~χ^2-dist w df n-2 & is indep from $\hat\beta_0$, $\hat\beta_1$ (verification given later)

Significant test: test whether or not there is a linear regression r/s btw the response var & the covariate

Null hypothesis, H_0: no such r/s. Alternative hypothesis, H_1: r/s exist

test statistic is F-statistic, where F = $\frac{MSR}{MSE}$ (intuitively, if variation caused by covariate > variation caused by error ⇒ r/s exist)

Under H_0, F~F-dist with df 1 and n-2, since SSR and SSE indep and each follows a χ^2-dist

If F > upper α quantile $f_{1, n-2}(\alpha)$, H_0 is rejected at level α, otherwise not rejected OR P(F > $f_{1, n-2}(\alpha)$) < α = P(type 1 error)

t-statistic for the inference on β_1

Dist of $\hat\beta_1$ cannot be used directly for inference, since its var involves σ^2 which is unknown

Let s²($\hat\beta_1$) = $\frac{s^2}{\sum_{i=1}^{n}(x_i-\bar{X})^2}$. Define $T_{\beta_1} = \frac{\hat\beta_1 - \beta_1}{s(\hat\beta_1)}$ = t-dist ~ t_{n-2}.

Statistical Inference for simple LRM

Two-sided test for β_1: H_0: β_1 = 0 and H_1: β_1 ≠ 0

Test-statistic: $T_{0\beta_1} = \frac{\hat\beta_1}{s(\hat\beta_1)}$ ~ t_{n-2} under H_0

Significance level α (usually taken as 0.05 or 0.01). If $|T_{0\beta_1}|$ > $t_{n-2}(\alpha/2)$, then reject H_0. Otherwise, don't reject H_0

OR p-value = p = 2P(|t_{n-2}| > |$T_{0\beta_1}$|) (T is observed value here, above was r.v). If p < α, reject H_0; otherwise, don't reject H_0

For simple LRM, the two-sided test for β_1 = p-value of F-statistic. F-statistic here = (t-statistic)²

One-sided test for β_1

1) H_0: β_1 ≤ 0, H_1: β_1 > 0. (Equal sign w H_0). If $T_{0\beta_1}$ > $t_{n-2}(\alpha)$, OR p = P(t_{n-2} > $T_{0\beta_1}$): reject H_0; otherwise don't

2) H_0: β_1 ≥ 0, H_1: β_1 < 0. If $T_{0\beta_1}$ < $t_{n-2}(\alpha)$, OR p = P(t_{n-2} < $T_{0\beta_1}$): reject H_0; otherwise don't

For both 2-sided, and 1-sided test, value 0 in hypotheses can be replaced by any constant c, then test statistic = $\frac{\hat\beta_1 - c}{s(\hat\beta_1)}$

From dist of T_{β_1}, 100(1 - α)% CI for β_1 is $[\hat{\beta}_1 - t_{n-2}(\alpha/2)s(\hat{\beta}_1), \hat{\beta}_1 + t_{n-2}(\alpha/2)s(\hat{\beta}_1)]$. CI are used for 2-sided tests

100(1- α)% lower confidence bound: $\beta_1 \geq \hat{\beta}_1 - t_{n-2}(\alpha)s(\hat{\beta}_1)$. This corresponds to 1-sided test (H_0: $\beta_1 \leq 0$, H_1: $\beta_1 > 0$)
100(1- α)% upper confidence bound: $\beta_1 \leq \hat{\beta}_1 + t_{n-2}(\alpha)s(\hat{\beta}_1)$. This corresponds to 1-sided test (H_0: $\beta_1 \geq 0$, H_1: $\beta_1 < 0$)

The inference on β_0
$T_{\beta_0} = \frac{\hat{\beta}_0-\beta_0}{s(\hat{\beta}_0)}$, where $s^2(\hat{\beta}_0) = \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n(x_i-\bar{X})^2}\right]\sigma^2$
Dist of T_{β_0} also t_{n-2}. Inference is similar to β_1. E.g. 95% CI for β_0 is $\hat{\beta}_0 \pm t_{n-2}(\alpha/2)s(\hat{\beta}_0)$

Prediction
CI of E(Y) when predictor value is x_h: $\hat{y}_h \pm t_{n-2}(\alpha/2)\sqrt{\mathrm{var}(\hat{Y})}$, where $\mathrm{var}(\hat{Y}) = \hat{\sigma}^2\left[\frac{1}{n} + \frac{(x_h-\bar{X})^2}{\sum_{i=1}^n(x_i-\bar{X})^2}\right]$ where $\hat{y}$ is the predicted value for x_h
Prediction interval of Y_{new} when predictor value is x_h: $\hat{y}_h \pm t_{n-2}(\alpha/2)\sqrt{var(Y_{new})}$, where $\mathrm{var}(Y_{new}) = \hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(X-\bar{X})^2}{\sum_{i=1}^n(x_i-\bar{X})^2}\right]$
$\mathrm{var}(Y_{new}) = \mathrm{var}(\hat{y}_h) + \hat{\sigma}^2$, where $\hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h$
Manual computation: $\mathrm{var}(\hat{y}_h) = \mathrm{var}(\hat{\beta}_0) + x_h^2\mathrm{var}(\hat{\beta}_1) + 2x_h\mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1)$. values and $\hat{\sigma}^2$ can be found from summary results

LSE as Method of moments estimation (MME): just replace theoretical moments w sample moments
From $\beta_1 = \rho_{xy}\frac{\sigma_y}{\sigma_x}$, $\beta_0 = \mu_y - \beta_1\mu_x \Rightarrow$ LSE is then $\hat{\beta}_1 = \hat{\rho}_{xy}\frac{\hat{\sigma}_y}{\hat{\sigma}_x}$ and $\hat{\beta}_0 = \hat{\mu}_y - \hat{\beta}_1\hat{\mu}_x$, where a qty w a hat = sample version of that qty

LSE as Maximum Likelihood estimation (MLE): Under the Normality assumption, the log-likelihood is
$-\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n(y_i - \beta_0 - \beta_1 x_i)^2 = -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}Q \Rightarrow$ MLE of σ^2 is given by $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^n(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \frac{n-2}{n}s^2$
Maximizing the log-likelihood to obtain the MLE of the regression coefficients is equivalent to minimizing Q

Multiple Linear Regression

Matrix form of linear fns: $\sum_{i=1}^n a_i x_i = a_1x_1 + \ldots + a_nx_n = \mathbf{a}^T\mathbf{x}$, where $\mathbf{a} = (a_1,\ldots,a_n)^T$, $\mathbf{x} = (x_1,\ldots,x_n)^T$
Matrix form of quadratic fns: $\sum_{i=1}^n\sum_{j=1}^n a_{ij}x_ix_j = \mathbf{x}^TA\mathbf{x}$, where $A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn}\end{pmatrix}$
Matrix form of differentiation: Let f($\mathbf{x}$) be a multivariate fn. Define $\frac{\partial f(x)}{\partial x} = \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n}\right)^T$,
then $\frac{\partial \boldsymbol{a}^T\mathbf{x}}{\partial x} = \mathbf{a}$, and $\frac{\partial x^T A x}{\partial x} = 2A\mathbf{x}$ if A is symmetric, $(A + A^T)\mathbf{x}$ otherwise

In multiple LRM, find r/s btw response var Y and p predictor variables/covariates $\mathbf{X} = (X_1,\ldots, X_p)$
Then multiple LRM is $Y = \beta_0 + \beta_1X_1 + \ldots + \beta_pX_p + \epsilon$, where ϵ is random error w mean 0
Then $EY = \beta_0 + \beta_1X_1 + \ldots + \beta_pX_p$ is the multiple linear regression fn. (Or more strictly speaking EY is E(Y|X))
For n observations, $y_i = \beta_0 + \beta_1x_{i1} + \ldots + \beta_px_{ip} + \epsilon_i$, i = 1,...,n

Let $X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np}\end{bmatrix}$, $\boldsymbol{\beta} = \begin{bmatrix}\beta_0 \\ \beta_1 \\ \vdots \\ \beta_p\end{bmatrix}$, $\mathbf{y} = \begin{bmatrix}y_0 \\ y_1 \\ \vdots \\ y_n\end{bmatrix}$, $\boldsymbol{\epsilon} = \begin{bmatrix}\epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_n\end{bmatrix}$. Then $\begin{bmatrix}y_0 \\ y_1 \\ \vdots \\ y_n\end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np}\end{bmatrix}\begin{bmatrix}\beta_0 \\ \beta_1 \\ \vdots \\ \beta_p\end{bmatrix} + \begin{bmatrix}\epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_n\end{bmatrix}$ OR $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$. X aka design matrix

Assumptions for multiple LRM (similar to simple LRM)	1. $X_1,\ldots,X_p$ and ϵ are indep.
2. ϵ_i have mean 0	3. ϵ_i are pairwise un-correlated, i.e. Cov(ϵ_i, ϵ_j) = 0
4. ϵ_i have common variance σ^2 (Homogeneity)	5. ϵ_i have a Normal dist (Normality)

Let μ_Y denote mean of Y, $\boldsymbol{\mu}_X$ denote mean vector of X = $(X_1,\ldots,X_p)$, Σ_{XY} : covariance vector btw Y and X, Σ_{XX}: covariance matrix of X and $\boldsymbol{\beta}_1 = (\beta_1,\ldots, \beta_p)$. By assumption1, $\boldsymbol{\beta}_1 = \Sigma_{XX}^{-1}\Sigma_{XY}$, $\beta_0 = \mu_Y - \boldsymbol{\beta}_1^T\boldsymbol{\mu}_X$

Least sq estimate for multiple LRM

Estimate $\beta_0, \beta_1, \ldots, \beta_p$ by minimizing $Q = \sum_{i=1}^n\left(y_i - \beta_0 - \beta_1x_{i1} - \ldots - \beta_px_{ip}\right)^2 = \left\|\mathbf{y} - X\boldsymbol{\beta}\right\|^2$ (norm)
LSE of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{y}$
$\hat{\sigma}^2 = \frac{SSE}{df} = \frac{\|\mathbf{y}-X\hat{\boldsymbol{\beta}}\|^2}{n-p-1} = \frac{\|\mathbf{y}-X(X^TX)^{-1}X^T\mathbf{y}\|^2}{n-p-1}$

Hat matrix & properties

$H = X(X^TX)^{-1}X^T$ is the hat matrix of X. It is the projection matrix of the linear space spanned by the cols of X
$H = H^T$, $HX = X$, $H^2 = H$, $(I-H)^2 = I-H$, $X^TH = X^T$
Residual vector: $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (I-H)\mathbf{y}$. Vector of fitted values: $\hat{\mathbf{y}} = H\mathbf{y}$
Hence $\mathbf{e}^T\mathbf{1} = 0$ (by assumption 2), $\mathbf{e}^T\mathbf{x}_j = 0$ (by assumption 1), mean of $\hat{\mathbf{y}} = \bar{\mathbf{y}}$

Decomposition of Sum of Squares

Matrices H_T, H_R and H_E are all symmetric and idempotent (i.e. $H^TH = H$)
Hence $SST = \mathbf{y}^TH_T\mathbf{y} = \mathbf{y}^T(I - \frac{\mathbf{11}^T}{n})\mathbf{y}$. $SSR = \mathbf{y}^TH_R\mathbf{y} = \mathbf{y}^T[X(X^TX)^{-1}X^T - \frac{\mathbf{11}^T}{n}]\mathbf{y}$. $SSE = \mathbf{y}^TH_E\mathbf{y} = \mathbf{y}^T[I - X(X^TX)^{-1}X^T]\mathbf{y}$. SST = SSR + SSE

Dist of Sum of Squares

Under assumptions of multiple regression models, (Note $\mathrm{var}(AY) = A\mathrm{var}(Y)A^T$, cov also similar)
Under hypothesis $\beta_1 = \ldots = \beta_p = 0$, SSR and SSE are indep. $\frac{SSE}{\sigma^2} \sim \chi^2_{n-p-1}$. $\frac{SSR}{\sigma^2} \sim \chi^2_p$,

ANOVA Table

Source of variation	SS	df	MS	F-statistic
Regression	SSR	p	MSR=SSR/p	MSR/MSE = $F_{p,\, n-p-1}$
Error	SSE	n-p-1	MSE=SSE/(n-p-1)	
Total	SST	n-1		

Coefficient of multiple determination

coefficient of multiple determination for multiple LRM is $R^2 = \frac{\sum_{i=1}^n(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^n(y_i-\bar{y})^2} = \frac{SSR}{SST} = \mathrm{corr}(\mathbf{y}, \hat{\mathbf{y}})^2$
Adjusted R^2: $R_a^2 = R^2 - \frac{p}{n-p-1}(1 - R^2)$

Multiple correlation coefficient

Correlation of response var Y w scalar covariate is measured by the Pearson's correlation coefficient
The multiple correlation coefficient btw response var Y and vector $\mathbf{z}$ of covariates $(X_1,\ldots,X_p)^T$ is $\mathrm{MCORR}(Y, \mathbf{z}) = \max_{\boldsymbol{a}}\mathrm{CORR}(Y, \boldsymbol{a}^T\boldsymbol{z})$ where $\mathbf{a}$ is a vector of constants, (i.e. $\mathbf{a}^T\mathbf{z}$ is LC of $\mathbf{z}$)
$[\mathrm{MCORR}(\mathrm{Y}, \mathbf{z})]^2 = \frac{\sum_{yz}\sum_{zz}^{-1}\sum_{zy}}{\sigma_y^2}$, where $\sum_{\boldsymbol{yz}} = (\mathrm{Cov}(Y, X_1),\ldots,\mathrm{Cov}(Y,X_p))$, $\sum_{\boldsymbol{zz}}$ = variance matrix of $\mathbf{z}$ and $\sum_{\boldsymbol{zy}} = \sum_{\boldsymbol{yz}}^T$, $\sigma_y^2 = \mathrm{var}(Y)$

Let $\mathbf{y}$ be vector of n observations of y, Z be the matrix of observed $\mathbf{z}$ w ith row given by $\mathbf{z}_i$.
$[\mathrm{MCORR}(\mathbf{y}, Z)]^2 = \frac{\mathbf{y}^T\left(I - \frac{\mathbf{11}^T}{n}\right)Z\left[Z^T\left(I - \frac{\mathbf{11}^T}{n}\right)Z\right]^{-1}Z^T\left(I - \frac{\mathbf{11}^T}{n}\right)\mathbf{y}}{\mathbf{y}^T\left(I - \frac{\mathbf{11}^T}{n}\right)\mathbf{y}} = \frac{cov(Z,y)^T[var(Z)]^{-1}cov(Z,y)}{var(y)}$
In multiple LRM w vector of p covariates $\mathbf{z} = (X_1,\ldots,X_p)^T$, $R^2 = [\mathrm{MCORR}(\mathbf{y}, Z)]^2$

Partial correlation coefficient	Let X(-j) denote the sub-matrix of the design matrix X obtained by deleting col $\mathbf{x}_j$ of X Let $\tilde{\boldsymbol{y}}$ be the residual of $\mathbf{y}$ regressed on X(-j), $\tilde{\boldsymbol{x}}_j$ is residual of $\mathbf{x}_j$ regressed on X(-j) Correlation btw $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{x}}_j$ is called partial correlation btw $\mathbf{y}$ and $\mathbf{x}_j$ adjusting for effects of X(-j), given by CORR$(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{x}}_j) = \frac{\sum_{i=1}^{n}(\tilde{y}_i - \bar{\tilde{y}})(\tilde{x}_{ij} - \bar{\tilde{x}_j})}{\sqrt{\sum_{i=1}^{n}(\tilde{y}_i - \bar{\tilde{y}})^2 \sum_{i=1}^{n}(\tilde{x}_{ij} - \bar{\tilde{x}_j})^2}} = \frac{\mathbf{y}^T(\mathrm{I} - H_{-j})\mathbf{y}}{\sqrt{\mathbf{y}^T(\mathrm{I} - H_{-j})\mathbf{y}\boldsymbol{x}_j^T(\mathrm{I} - H_{-j})\boldsymbol{x}_j}}$ where H_{-j} is the hat matrix of X(-j) Since residual of $\mathbf{y}$ regressed on X(-j) is part of $\mathbf{y}$ unexplained by X(-j), the squared partial correlation btw $\mathbf{y}$ and $\mathbf{x}_j$ adjusting for effects of X(-j) is the proportion of unexplained variation of $\mathbf{y}$ which is explained by $\mathbf{x}_j$
Explicit expression of $\hat{\beta}_j$	Explicit expression of $\hat{\beta}_j$ can be obtained by considering the minimization of $\lVert \boldsymbol{y} - X\boldsymbol{\beta} \rVert^2$: 1st) minimize wrt $\boldsymbol{\beta}_{-j}$ with β_j fixed, then minimize wrt β_j, where $\boldsymbol{\beta}_{-j}$ is the sub-vector of $\boldsymbol{\beta}$ eliminating β_j
	Minimizing w fixed β_j: $\min_{\boldsymbol{\beta}_{-j}} \lVert \boldsymbol{y} - X\boldsymbol{\beta} \rVert^2 = \min_{\boldsymbol{\beta}_{-j}} \lVert (I - H_{-j})\boldsymbol{y} - \beta_j (I - H_{-j})\boldsymbol{x}_j \rVert^2$
	Minimizing wrt β_j: $\hat{\beta}_j = \frac{\boldsymbol{x}_j^T(I - H_{-j})\boldsymbol{y}}{\boldsymbol{x}_j^T(I - H_{-j})\boldsymbol{x}_j}$; var$(\hat{\beta}_j) = \frac{\sigma^2}{\boldsymbol{x}_j^T(I - H_{-j})\boldsymbol{x}_j}$
	$\frac{\hat{\beta}_j}{sd(\hat{\beta}_j)}$ (Standardization) = $\frac{\boldsymbol{x}_j^T(I - H_{-j})\boldsymbol{y}}{\sigma\sqrt{\boldsymbol{x}_j^T(I - H_{-j})\boldsymbol{x}_j}}$, (which tells us which covariate contribute more to variation in y) Note R_j^2 = CORR$(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{x}}_j) = \frac{\mathbf{y}^T(\mathrm{I} - H_{-j})\mathbf{y}}{\sqrt{\mathbf{y}^T(\mathrm{I} - H_{-j})\mathbf{y}}\sqrt{\boldsymbol{x}_j^T(\mathrm{I} - H_{-j})\boldsymbol{x}_j}}$ = some constant * $\frac{\hat{\beta}_j}{sd(\hat{\beta}_j)}$
Example	In ANOVA table, the sum of squares (SS) generated are sequential SS, ie. the SS associated w each covariate is the one after adjusting for the effects of the covariate preceding it. SSR of y~x1+x2 is sum of all sequential SS
Properties of LSE	Dist of $\hat{\beta}$ is normal w mean = $\boldsymbol{\beta}$, var = $\sigma^2(X^TX)^{-1}$
	(n-p-1)$\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}$; $\hat{\sigma}^2$ and $\widehat{\boldsymbol{\beta}}$ are indep
	Let $\widehat{\Sigma} = \hat{\sigma}^2(X^TX)^{-1}$. Let $\hat{\sigma}_j^2$ denote the jth diagonal elems of $\widehat{\Sigma}$ (estimated var of $\hat{\beta}_j$) Note $\hat{\sigma}_j^2 = c_{jj}\hat{\sigma}^2$, where c_{jj} is the jth diagonal elem of $(X^TX)^{-1}$ Then $\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j^2} \sim t_{n-p-1}$; & For any constant vector $\mathbf{c}$, $\frac{\boldsymbol{c}^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sqrt{\boldsymbol{c}^T\widehat{\Sigma}\boldsymbol{c}}} \sim t_{n-p-1}$.
	MSR and MSE are indep, since SSR and SSE are indep
Significance F-test	Hypothesis: H_0: $\beta_1 = \ldots = \beta_p = 0$ vs H_1: $\beta_j \neq 0$ for at least one of j = 1,...,p. Test statistic: F = MSR/MSE. Under H_0, F ~ $F_{p,\, n-p-1}$ For a significance level α, reject H_0 if F ≥ $f_{p,n-p-1}(\alpha)$ or the p-value P($F_{p,n-p-1}$ ≥ F) < α; otherwise, do not reject H_0
Wald test statistic	Let $\boldsymbol{\theta}$ be a vector of parameters, $\widehat{\boldsymbol{\theta}}$ be its estimator, and $\widehat{\Sigma}_\theta$ the estimated var matrix of $\widehat{\boldsymbol{\theta}}$ Wald statistic for testing H_0: $\boldsymbol{\theta}$ = 0 is given by $\widehat{\boldsymbol{\theta}}^T\widehat{\Sigma}_\theta^{-1}\widehat{\boldsymbol{\theta}}$ Thus Wald test statistic for significance test H_0: $\boldsymbol{\beta}_1$ = 0, where $\boldsymbol{\beta}_1 = (\beta_1,\ldots, \beta_p)^T$ is given by W = $\widehat{\boldsymbol{\beta}}_1^T\widehat{\Sigma}_\beta^{-1}\widehat{\boldsymbol{\beta}}_1$, where $\widehat{\boldsymbol{\beta}}_1$ is estimator of $\boldsymbol{\beta}_1$ and $\widehat{\Sigma}_\beta^{-1}$ is estimated variance matrix of $\widehat{\boldsymbol{\beta}}_1$ In context of multiple LRM, F = W/p, where p is dimension of $\boldsymbol{\beta}_1$
Individual t-test	Answers qn: given other vars in model, does a particular predictor have a significant effect? Hypotheses: H_0 : β_j = 0 vs H_1: $\beta_j \neq 0$. Test statistic: T = $\frac{\widehat{\beta}_j}{\hat{\sigma}_j}$, where $\hat{\sigma}_j$ is the estimated SD of $\hat{\beta}_j$. Under H_0, T ~ t_{n-p-1} For a significance level α, reject H_0 if |T| ≥ $t_{n-p-1}(\alpha/2)$ or p-value 2P(t_{n-p-1} ≥ |T|) < α; otherwise do not reject H_0 - p-value better than test statistic as not only can reject H_0, if p-value very small -> evidence supporting H_1 is strong - 1-sided test also same way as in simple LRM
Testing general linear hypothesis	General linear hypothesis : H_0 : $\sum_{j=0}^{p} c_j\beta_j = 0$ For testing linear hypothesis, test statistic is T = $\frac{\boldsymbol{c}^T\widehat{\boldsymbol{\beta}}}{\sqrt{\boldsymbol{c}^T\widehat{\Sigma}\boldsymbol{c}}}$. Under H_0, t ~ t_{n-p-1} If only a few components of $\mathbf{c}$ are non-zero, T can be simplified. E.g. $\mathbf{c} = (c_1,c_2,0,\ldots,0)^T$, the var $\boldsymbol{c}^T\widehat{\Sigma}\boldsymbol{c}$ becomes c_1^2var$(\hat{\beta}_1)$ + c_2^2var$(\hat{\beta}_2)$ + $2c_1c_2$cov$(\hat{\beta}_1, \hat{\beta}_2)$ and $\boldsymbol{c}^T\widehat{\boldsymbol{\beta}}$ becomes $c_1\hat{\beta}_1 + c_2\hat{\beta}_2$ For a significance level α, reject H_0 if |T| ≥ $t_{n-p-1}(\alpha/2)$ or p-value 2P(t_{n-p-1} ≥ |T|) < α; otherwise do not reject H_0
CI and confidence bound	A 100(1 - α)% CI for β_j is $[\hat{\beta}_j - \hat{\sigma}_j t_{n-p-1}(\alpha/2), \hat{\beta}_j + \hat{\sigma}_j t_{n-p-1}(\alpha/2)]$ A 100(1 - α)% CI for $\boldsymbol{c}^T\boldsymbol{\beta}$ is $\left[\boldsymbol{c}^T\widehat{\boldsymbol{\beta}} - \sqrt{\boldsymbol{c}^T\widehat{\Sigma}\boldsymbol{c}} * t_{n-p-1}(\alpha/2), \boldsymbol{c}^T\widehat{\boldsymbol{\beta}} + \sqrt{\boldsymbol{c}^T\widehat{\Sigma}\boldsymbol{c}} * t_{n-p-1}(\alpha/2)\right]$ The 100(1 - α)% confidence bounds for β_j are $\begin{cases}\text{Upper bound: } \beta_j \leq \hat{\beta}_j + \hat{\sigma}_j t_{n-p-1}(\alpha) \\ \text{Lower bound: } \beta_j \geq \hat{\beta}_j - \hat{\sigma}_j t_{n-p-1}(\alpha)\end{cases}$
Prediction	Given a new observation $\mathbf{x}_0 = (1, x_{01},\ldots, x_{0p})^T$, predicted value for both Ey_0 and y_0 is $\hat{y}_0 = \boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}}$ Estimated variance of fitted value is $\hat{\sigma}_F^2(\hat{y}_0) = \boldsymbol{x}_0^T\hat{\sigma}^2(X^TX)^{-1}\boldsymbol{x}_0$ Estimated prediction error variance is $\hat{\sigma}_P^2(\hat{y}_0) = \hat{\sigma}^2 + \hat{\sigma}_F^2(\hat{y}_0)$ Note $y_0 = \boldsymbol{x}_0^T\boldsymbol{\beta} + \epsilon$. But $Ey_0 = \boldsymbol{x}_0^T\boldsymbol{\beta}$. Thats why prediction have extra error term CI for Ey_0 is $\hat{y}_0 \pm \hat{\sigma}_F^2(\hat{y}_0)t_{n-p-1}(\alpha/2)$ Prediction interval for y_0 is $\hat{y}_0 \pm \hat{\sigma}_P^2(\hat{y}_0)t_{n-p-1}(\alpha/2)$

One-Way ANOVA	One-Way ANOVA model deals w only one factor, A having a levels. $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, i = 1,...,a, j=1,...,$n_i$, where y_{ij} is the value of Y for the j-th member of the i-th grp, μ and α_i are unknown params, ϵ_{ij} are iid random errors, and n_i is num of observations in i-th grp Let μ_i be mean response to i-th grp. $\mu_i = \mu + \alpha_i$. For identifiability, we restrict $\sum_{i=1}^{a}\alpha_i = 0$. Then μ represents the overall mean and α_i represents effect of i-th treatment
	Answers qn whether factor A has any effect (i.e. any diff btw its levels) by analyzing the components of the variation in Y SST = SSA + SSE. Total sum of squares = sum of squares attributable to factor A and sum of squares attributable to errors SST = $\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$, SSA = $\sum_{i=1}^{a} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$, SSE = $\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$

$\bar{y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$, $\bar{y}_{..} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}$, n = $\sum_{i=1}^{a} n_i$

One-way ANOVA table is:

Source	df	SS	MS	F-test
A	a-1	SSA	MSA	MSA/MSE
Error	n-a	SSE	MSE	
Total	n-1	SST		

Effect of the factor is tested by the F-statistic, F = MSA/MSE

Under hypothesis α_1 = ... = α_a = 0, F~$F_{a-1,n-a}$. p-value for F-test use pf(F-statistic, a-1, n-a, lower.tail=FALSE)

Two-way ANOVA

Allows analysis of 2 factors at the same time & analysis of interaction of 2 factors

Analyse effect of 2 factors A and B on response variable Y. Suppose A has a levels, B has b levels

$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$, i = 1,...,a, j = 1,...,b, k = 1,...,n_{ij}, (gamma: interaction btw the 2 factors) where

$\sum_{i=1}^{a}\alpha_i$ = 0, $\sum_{j=1}^{b}\beta_j$ = 0, $\sum_{i=1}^{a}\gamma_{ij}$ = $\sum_{j=1}^{b}\gamma_{ij}$ = 0

Two-way ANOVA investigates whether there is interaction btw the 2 factors, whether the main effect (ave effects of 1 factor over all the levels of the other factor) is significant

SST = SSA + SSB + SSE + SSAB (sum of squares due to interaction of factor A and B)

SST = $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{...})^2$	SSA = $\sum_{i=1}^{a} n_{i.}(\bar{y}_{i..} - \bar{y}_{...})^2$	SSB = $\sum_{j=1}^{b} n_{.j}(\bar{y}_{.j.} - \bar{y}_{...})^2$	SSE = $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij.})^2$
SSAB = $\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$	$\bar{y}_{i..} = \frac{1}{n_{i.}}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk}$	$n_{i.} = \sum_{j=1}^{b} n_{ij}$	$\bar{y}_{.j.} = \frac{1}{n_{.j}}\sum_{i=1}^{a}\sum_{k=1}^{n_{ij}} y_{ijk}$

$\bar{y}_{.j.} = \frac{1}{n_{.j}}\sum_{i=1}^{a}\sum_{k=1}^{n_{ij}} y_{ijk}$	$n_{.j} = \sum_{i=1}^{a} n_{ij}$	$\bar{y}_{...} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk}$,	n = $\sum_{i=1}^{a} n_{i.}$	$\bar{y}_{ij.} = \frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}} y_{ijk}$

Source	df	SS	MS	F-test
A	a-1	SSA	MSA = SSA/(a-1)	MSA/MSE
B	b-1	SSB	MSB = SSB/(b-1)	MSB/MSE
AB	(a-1)(b-1)	SSAB	MSAB = SSAB/(a-1)(b-1)	MSAB/MSE
Error	n-ab	SSE	MSE = SSE/(n-ab)	
Total	n-1	SST		

Test for interaction: F = MSAB/MSE. Under hypothesis of no interaction F~$F_{(a-1)(b-1), n-ab}$

Test for main effect: F_1 = MSA/MSE, F_2 = MSB/MSE. Under hypothesis of zero main effects, F_1~$F_{a-1, n-ab}$, F_2~$F_{b-1, n-ab}$

Factor w zero main effect $\nRightarrow$ factor has no effect, unless the interaction effect DNE

p-value for F_1 = pf(F_1, a-1, n-ab, lower.tail=FALSE). p-value for F_2 = pf(F_2, b-1, n-ab, lower.tail=FALSE)

p-value for F = pf(F, (a-1)(b-1), n-ab, lower.tail=FALSE)

A and B have a significant interaction effect on Y; the main effect of A is significant, but main effect of B is not

Main effect & Contrasts

If change in 1 var elicits a change in another var, then the var has an effect on the other var

Effect of a factor is the differences of the expected response it causes among its levels

Effect can be measure by contrasts. Let μ_k denote expected response at level k.

Contrast is defined as $\sum_{k=1}^{a} c_k\mu_k$, w $\sum_{k=1}^{a} c_k$ = 0, i.e. $\mathbf{c}^T\mathbf{1}$ = 0, where **c** = (c_1,...,c_k) is called a contrast vector

If factor has no effect, then all contrasts are 0, i.e. $\mu_1 = \mu_2 = \ldots = \mu_a \Leftrightarrow \sum_{k=1}^{a} c_k\mu_k = 0$, for all **c**

Interaction Effect

If effect of factor A is diff when factor B is fixed at diff levels, or equivalently, effect of B is diff when A is fixed at diff levels, then it is said that the 2 factors have an interaction effect

E.g. If A and B only have 2 levels, the effects of A at the 2 levels of B are: $\mu_{21} - \mu_{11}$, $\mu_{22} - \mu_{12}$.

If both not same -> there is interaction btw A and B.

Interaction effect is measured by $(\mu_{22} - \mu_{12}) - (\mu_{21} - \mu_{11}) = \mu_{22} - \mu_{12} - \mu_{21} + \mu_{11}$

In general, effect of factor A at a fixed level, i, of B is measured by any contrasts $\sum_{k=1}^{a} c_k\mu_{ki}$. If there is at least one pair (i,j) and at least one contrast **c**, s.t. $\sum_{k=1}^{a} c_k\mu_{ki} \neq \sum_{k=1}^{a} c_k\mu_{kj}$, then A and B have interaction effect

Interaction effect is measured by interaction contrasts: (i.e. contrast of the contrast)

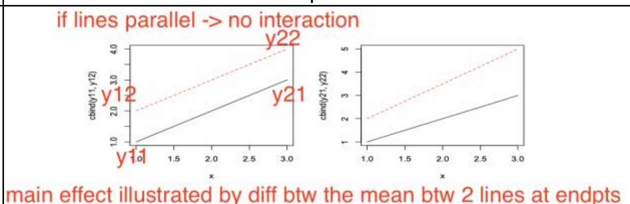
$\sum_{j=1}^{b} d_j\left[\sum_{k=1}^{a} c_k\mu_{kj}\right] = \sum_{j=1}^{b}\sum_{k=1}^{a} c_k d_j\mu_{kj}$, where $\sum_{k=1}^{a} c_k = 0$, $\sum_{j=1}^{b} d_j = 0$

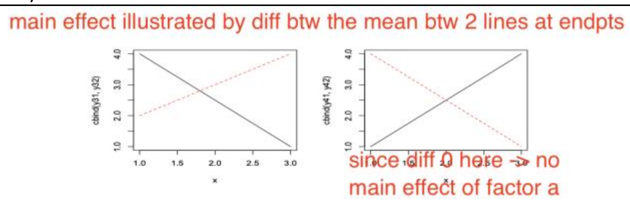
Interaction contrast vector is of the form (c_1d_1,...,c_1d_b,...,c_ad_1,...,c_ad_b)

Components of the vector have restrictions: $\sum_{k=1}^{a} c_k d_j = 0$, j=1,...,b; $\sum_{j=1}^{b} c_k d_j = 0$, k=1,...,b; $\sum_{k=1}^{a}\sum_{j=1}^{b} c_k d_j = 0$

Among the 1st b restrictions, only b-1 are indep, among 2nd a restrictions, only a-1 are indep, and altogether there are a+b-1 indep restrictions. Thus num of indep interaction contrasts is ab-a-b+1 = (a-1)(b-1)

Illustration of main and interaction effects





Limitation of ANOVA

Not convenient for detailed analysis of effects

In two-way ANOVA model, if n_{ij} (group sizes) are not the same, the SSAB does not measure the interaction effect, i.e. in SSAB, the term $\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$ is not the estimate of an interaction contrast

ANOVA by LRM

Factor predictor can be represented by dummy vars. If factor has a levels, it can be represented by a-1 dummy vars

Dummy vars: $u_k = \begin{cases} 1, \text{ if level } k \\ 0, \text{otherwise} \end{cases}, k = 2, \ldots, a$

LRM for one-way ANOVA

Model: $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, expressed in dummy variables u_k: $y_{ij} = \beta_0 + \beta_2 u_2 + \ldots + \beta_a u_a + \epsilon_{ij}$

In matrix form: **y** = Xβ + $\boldsymbol{\epsilon}$, where **y** = $(y_{11}, \ldots, y_{1n_1}, \ldots, y_{a1}, \ldots, y_{an_a})^T$, $\boldsymbol{\epsilon}$ = $(\epsilon_{11}, \ldots, \epsilon_{1n_1}, \ldots, \epsilon_{a1}, \ldots, \epsilon_{an_a})^T$, X = $\begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{1}_{n_a} & \mathbf{0} & \cdots & \mathbf{1}_{n_a} \end{pmatrix}$.

Alternatively, model can be expressed as $y_{ij} = \begin{cases} \beta_0 + \epsilon_{1j}, \text{ if level } 1\ (i=1) \\ \beta_0 + \beta_k + \epsilon_{kj}, \text{if level } k\ (i=k), k = 2, \ldots, a \end{cases}$

Regression params and mean response at ea level have relation: $\mu_1 = \beta_0$, $\mu_k = \beta_0 + \beta_k$, k=2,...,a

	Hence param $\beta_k = \mu_k - \mu_1$ (which is a contrast) is the diff btw expected values at level k annd level1
	LRM for two-way ANOVA 2 cases: (i) no repeated observations at ea level combination, i.e. $n_{ij} = 1$; (ii) $n_{ij} > 1$ (i): only main effects can be analyzed; (ii) both main effects and interaction effects can be analyzed
	Main effect models: Dummy var for 2 factors: $u_i = \begin{cases} 1, \ if\ level\ i \\ 0, otherwise \end{cases}, i = 2, \dots, a,\ v_j = \begin{cases} 1, \ if\ level\ j \\ 0, otherwise \end{cases}, j = 2, \dots, b$ Main effect model: $y_{ijk} = \beta_0 + \sum_{i=2}^{a} \alpha_i u_i + \sum_{j=2}^{b} \beta_j v_j + \epsilon_{ijk}$ Expectation of the main effect model at diff levels: $Ey_{ijk} = \begin{cases} \mu_{11} = \beta_0, & both\ A\ and\ B\ at\ level\ 1 \\ \mu_{i1} = \beta_0 + \alpha_i, & A\ at\ level\ i, B\ at\ level\ 1 \\ \mu_{1j} = \beta_0 + \beta_j, & A\ at\ level\ 1, B\ at\ level\ j \\ \mu_{ij} = \beta_0 + \alpha_i + \beta_j, & A\ at\ level\ i, B\ at\ level\ j \end{cases}$ $\alpha_i = u_{i1} - u_{11}$: diff of effect of A at level i and 1 when B is fixed at level 1, which is the same as $u_{ij} - u_{1j}$, the diff of effects of A at level k and 1 when B is fixed at level j. i.e. diff does not depend on j $\beta_j = u_{1j} - u_{11}$, diff of effects of B at level j and 1 when A is fixed at level 1, which is the same as $u_{ij} - u_{i1}$, the diff of effects of B at level j and 1 when A is fixed at level i. i.e. diff does not depend on i
	Interaction model: $y_{ijk} = \beta_0 + \sum_{i=2}^{a} \alpha_i u_i + \sum_{j=2}^{b} \beta_j v_j + \sum_{i=2}^{a} \sum_{j=2}^{b} \gamma_{ij} u_i v_j + \epsilon_{ijk}$ Expectation of the main effect model at diff levels: $Ey_{ijk} = \begin{cases} \mu_{11} = \beta_0, & both\ A\ and\ B\ at\ level\ 1 \\ \mu_{i1} = \beta_0 + \alpha_i, & A\ at\ level\ i, B\ at\ level\ 1 \\ \mu_{1j} = \beta_0 + \beta_j, & A\ at\ level\ 1, B\ at\ level\ j \\ \mu_{ij} = \beta_0 + \alpha_i + \beta_j + \xi_{ij}, & A\ at\ level\ i, B\ at\ level\ j \end{cases}$ $\alpha_i = \mu_{i1} - \mu_{11}$: diff of effect of A at level i and 1 when B is fixed at level 1, $\beta_j = \mu_{1j} - \mu_{11}$, diff of effects of B at level j and 1 when A is fixed at level 1, $\alpha_i + \xi_{ij} = \mu_{ij} - \mu_{1j}$: diff of effect of A at level i and 1 when B is fixed at level j, $\beta_j + \xi_{ij} = \mu_{ij} - \mu_{i1}$, diff of effects of B at level j and 1 when A is fixed at level i, $\xi_{ij} = (\mu_{ij} - \mu_{i1}) - (\mu_{1j} - \mu_{11}) = \gamma_{ij}$, interaction contrast (diff of effect of A at level i and 1 when B is at level j and 1) So main effect caused by level i and level 1 of A is the average over all the levels of B, i.e. $\alpha_i + \frac{1}{b}\sum_{j=2}^{b} \xi_{ij}$
Inference on interaction effect	For testing whether there is a significant overall interaction effect, under the regression model, H_0: $\xi_{ij} = 0$, i=2,...,a; j=2,...,b; vs H_1: at least one of $\xi_{ij} \neq 0$, where ξ_{ij} are the basis interaction contrasts Use F-test statistic to test hypothesis 1) F-test statistic from table produced by anova 2) F-test statistic from computation of Wald statistic: W = $\hat{\boldsymbol{\xi}}^T \widehat{\Sigma}_{\xi}^{-1} \hat{\boldsymbol{\xi}}$, where $\hat{\boldsymbol{\xi}}$ is vector of estimated ξ_{ij}'s and $\widehat{\Sigma}_{\xi}$ is the estimated covariance matrix of $\hat{\boldsymbol{\xi}}$. F-statistic is F = $\frac{W}{(a-1)(b-1)}$ Under H_0 $F \sim F_{(a-1)(b-1),\ n-ab}$. Reject H_0 at level α, if $F > F_{(a-1)(b-1),\ n-ab}$ (α) or p-value pf(F, (a-1)(b-1), n-ab, lower.tail=FALSE) < α
	For testing of particular interaction effect: perform individual test on ξ_{ij}'s and on linear combination of ξ_{ij}'s E.g. 1) Test whether diff btw level j and 1 of Factor B is same at level i and level 1 of Factor A, i.e. $H_0 : (u_{ij} - u_{i1}) - (u_{1j} - u_{11}) = \xi_{ij} = 0$ 2) Test whether diff btw level j and l of Factor B is same at level i and level 1 of Factor A, i.e. $H_0 : (u_{ij} - u_{il}) - (u_{1j} - u_{1l}) = \xi_{ij} - \xi_{il} = 0$ 3) Test whether diff btw level j and k of B is same at level i and l of Factor A, i.e. $H_0 : (u_{ij} - u_{ik}) - (u_{lj} - u_{lk}) = \xi_{ij} - \xi_{ik} - \xi_{lj} + \xi_{lk} = 0$ General rule of thumb: if subscript of u contains 1: ignore; else convert to ξ Since any particular interaction contrast is a LC of ξ_{ij}'s, i.e. $\mathbf{c}^T\boldsymbol{\xi}$ where $\boldsymbol{\xi}$ is the vector of ξ_{ij}. Let $\hat{\boldsymbol{\xi}}$ be the vector of estimated ξ_{ij}'s and $\widehat{\Sigma}_{\xi}$ is the estimated covariance matrix of $\hat{\boldsymbol{\xi}}$ Test statistic for $\mathbf{c}^T\boldsymbol{\xi} = 0$ is $T_{\mathbf{c}} = \frac{\mathbf{c}^T \hat{\boldsymbol{\xi}}}{\sqrt{\mathbf{c}^T \widehat{\Sigma}_{\xi} \mathbf{c}}}$. Under H_0: $\mathbf{c}^T\boldsymbol{\xi} = 0$, $T_{\mathbf{c}} \sim t_{n-ab}$ E.g. for a particular interaction contrast, above formula can be simplified. H_0: $\xi_{ij} - \xi_{ik} - \xi_{lj} + \xi_{lk} = 0$. Only vector $\widehat{\boldsymbol{\theta}} = \left(\hat{\xi}_{ij}, \hat{\xi}_{ik}, \hat{\xi}_{lj}, \hat{\xi}_{lk},\right)^T$ and its covariance matrix $\widehat{\Sigma}_{\theta}$ are needed. And corresponding $\mathbf{c}$ reduces to $\mathbf{b} = (1, -1, -1, 1)^T$
Example main effect contrast	Main effect contrasts cannot be conveniently calculated using iteraction model Main effect model has correct estimates for main-effect contrasts. But estimate $\hat{\sigma}_M^2$ from main-effect model not correct estimate of error variance. $SSE_M = SSE_I + SSAB$ (error caused by iteraction + interaction effect). Instead, $\hat{\sigma}_I^2$ from iteraction model has correct variance
	Let X_m denote design matrix of main-effect model. The estimated var matrix from main-effect model is V.m = $\hat{\sigma}_M^2 (X_m^T X_m)^{-1}$ It shld be adjusted to V = $\hat{\sigma}_I^2 (X_m^T X_m)^{-1} = (\hat{\sigma}_I^2 / \hat{\sigma}_M^2)$V.m Test statistic for main-effect contrast can be computed using coefficient from main-effect model and the adjusted estimated var matrix
Remarks	Insignificance of the main effect does not imply it has no effect if its iteraction w another factor is significant When interaction is significant, levels of a factor shld be compared at ea level of the other factor In general, inference on main effect when iteraction is significant is not very relevant

Family-wise type I error rate	In multiple comparison problem, we investigate many contrasts which form a family. If a contrast is "actually" 0 but turn out to be "significant", these apparently significant contrasts are called artifacts To avoid claiming artifacts as significant contrasts, the family-wise type I error rate must be controlled Denote family of contrasts by $\mathcal{C}$. For any $j \in \mathcal{C}$, let T_j be test statistic & decision rule to reject H_0 be $T_j \geq c_j$ for a critical value c_j Family-wise type I error rate is $P(\cup_{j \in \mathcal{C}} \{T_j \geq c_j\})$
General exploration	Find if there is any contrasts of grp means which are statistically significant. For main effects: grp means are those at the level of a factor For interaction effect: grp means are those at the level combination of 2 factor So let $\boldsymbol{v}$ be the vector of the grp means. Let $\mathbf{c}$ denot a contrast vector of $\boldsymbol{v}$ and $\mathcal{C}$ denote set of all possible contrasts. Then H_0: $\mathbf{c}^T\boldsymbol{v} = 0$ for all $\mathbf{c} \in \mathcal{C}$. H_0 can be expressed in terms of the basis contrasts: Let $\boldsymbol{\theta}$ be the vector of basis contrasts. H_0 is then equivalent to $\mathbf{a}^T\boldsymbol{\theta} = 0$ for any $\mathbf{a}$ In main effects: $\boldsymbol{\theta} = (\alpha_2, \dots, \alpha_a)$ or $(\beta_2, \dots, \beta_b)$. In interaction effects: $\boldsymbol{\theta} = (\xi_{ij})_{i=2,\dots,a;\ j=2,\dots,b}$
	Scheffe's sol^n For an individual contrast $\mathbf{a}^T\boldsymbol{\theta}$, test statistic $T_{\mathbf{a}} = \frac{\mathbf{a}^T \widehat{\boldsymbol{\theta}}}{\sqrt{\mathbf{a}^T \widehat{\Sigma}_{\theta} \mathbf{a}}}$, where $\widehat{\boldsymbol{\theta}}$ is estimate of $\boldsymbol{\theta}$, $\widehat{\Sigma}_{\theta}$ is estimated var matrix of $\widehat{\boldsymbol{\theta}}$

	Need to find critical value c_α s.t. P($\cup_{all\ \boldsymbol{a}}\{\lvert T_{\boldsymbol{a}}\rvert \geq c_\alpha\}\mid H_0$) ≤ α To control family-wise type I error rate at level α, Scheffe's soln is to take for all j, c_j = c_α = $\sqrt{m_1 F_{m_1,m_2}(\alpha)}$, where m_1 is num of components of $\boldsymbol{\theta}$ and m_2 is df of $\hat{\sigma}^2$, $F_{m_1,m_2}(\alpha)$ is the upper α quantile of the F_{m_1,m_2} dist. (c_α aka Scheffe's criterion)
One-way ANOVA	If multiple comparison is done directly w contrasts of level means (and not in terms of indep contrasts, i.e. Scheffe's), the test statistic for contrast $\sum_{j=1}^{a} c_j\mu_j$ is $T_{\boldsymbol{c}} = \frac{\sum_{k=1}^{a} c_k\bar{Y}_k}{\sqrt{\hat{\sigma}^2\sum_{k=1}^{a}\frac{c_k^2}{n_k}}}$ and $\max_{\boldsymbol{c}\in\mathcal{R}^a} T_{\boldsymbol{c}}^2 = \frac{1}{\hat{\sigma}^2}\max_{\boldsymbol{c}\in\mathcal{R}^a}\frac{\lvert\sum_{k=1}^{a} c_k\bar{Y}_k\rvert^2}{\sum_{k=1}^{a}\frac{c_k^2}{n_k}}$. Note $\sum_{k=1}^{a} c_k\bar{Y}_k = \sum_{k=1}^{a} c_k(\bar{Y}_k - \bar{Y})$ Let **b** = $(\bar{Y}_1 - \bar{Y}, \ldots, \bar{Y}_a - \bar{Y})^T$, **c** = $(c_1,\ldots,c_a)^T$, D = diag($n_1^{-1}, \ldots, n_a^{-1}$). Then $\max_{\boldsymbol{c}\in\mathcal{R}^a}\frac{\lvert\sum_{k=1}^{a} c_k\bar{Y}_k\rvert^2}{\sum_{k=1}^{a}\frac{c_k^2}{n_k}} = \max_{\boldsymbol{c}\in\mathcal{R}^a}\frac{\boldsymbol{c}^T\boldsymbol{b}\boldsymbol{b}^T\boldsymbol{c}}{\boldsymbol{c}^T D\boldsymbol{c}}$ = $\mathbf{b}^T D^{-1}\mathbf{b}$ = SSA = $\sum_{i=1}^{a} n_i(\bar{y}_i - \bar{y})^2$ And $\max_{\boldsymbol{c}\in\mathcal{R}^a} T_{\boldsymbol{c}}^2$ = (a-1)MSA/MSE = SSA/MSE (since SSA = (a-1)MSA, $\hat{\sigma}^2$ = MSE) (where MSA/MSE is $F_{a-1,n_{ERR}}$) p-value = P($F_{a-1,n-a} \geq$ F-ratio)
	Then the simultaneous p-value of 2-sided test is P($\cup_{all\ \boldsymbol{a}}\{\lvert T_{\boldsymbol{a}}\rvert \geq c_\alpha\}$) = P($\max_{\boldsymbol{c}\in\mathcal{R}^a} T_{\boldsymbol{c}}^2 \geq c_\alpha$) = P($F_{a-1,n-a} \geq T_{\boldsymbol{c}}^2/(a-1)$) For 1-sided test: p-value is (1/2)P($F_{a-1,n-a} \geq T_{\boldsymbol{c}}^2/(a-1)$) and critical value is $c_{2\alpha} = \sqrt{(a-1)F_{a-1,n-a}(2\alpha)}$,
Individual & Simultaneous CI	Simultaneous CI: CI which covers all params Simultaneous CI w confidence coefficient 1-α for all j require that P($\cap_{j=1}^{m}\{L_j \leq \theta_j \leq U_j\}$) ≥ 1-$\alpha$ Simulateous CI for the contrasts $\sum_{k=1}^{a} c_k\mu_k$ is $\sum_{k=1}^{a} c_k\bar{Y}_k \pm \sqrt{\hat{\sigma}^2\sum_{k=1}^{a}\frac{c_k^2}{n_k}}\sqrt{(a-1)F_{a-1,n_{ERR}}(\alpha)}$
Approach for general exploring	1) Conduct overall significance test to see if there is any effect on Y (if no -> stop) 2) If significance test is significant, find particular significant effects (impossible to investigate all possible contrasts, so just look at rather diff $\bar{Y}_k$ in summary data or look at estimated regression coefficients)
Multiple comparison through LRM	LRM for one-way ANOVA is $y_i = \beta_0 + \beta_2 u_{i2} + \ldots + \beta_a u_{ia} + \epsilon_i$, i = 1,..,n where $\beta_k = \mu_k - \mu_1$, k ≥ 2 For any contrast, we have $\sum_{k=1}^{a} c_k\mu_k = \sum_{k=2}^{a} c_k\mu_k - \sum_{k=2}^{a} c_k\mu_1 = \sum_{k=2}^{a} c_k(\mu_k - \mu_1) = \sum_{k=2}^{a} c_k\beta_k$ Test statistic: $T_C = \frac{\sum_{k=1}^{a} c_k\hat{\beta}_k}{\sqrt{\tilde{\boldsymbol{c}}^T\hat{\Sigma}_\beta\tilde{\boldsymbol{c}}}}$, where $\tilde{\boldsymbol{c}}$ = $(c_2,\ldots,c_a)^T$, $\hat{\Sigma}_\beta$ is estimated covariance matrix of ($\hat{\beta}_2,\ldots,\hat{\beta}_a$) T_C is same as T_C when using level means
Pairwise contrasts	Only interested in certain pairwise contrasts, $\mu_k - \mu_j$, 1 ≤ k < j ≤ a. Overall significance F-test not necessary Only need to find c_α s.t P$\left(\max_{k,j}\frac{\lvert\bar{Y}_k - \bar{Y}_j\rvert}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n_k}+\frac{1}{n_k}\right)}} \geq c_\alpha\right)$ = α
Studentized range dist	Let $\bar{Y}_i$, i=1,...,a be sample means of a samples w equal sizes (n_i = n) Studentized range statistic is $q_{a,n_{ERR}}$ = $\sqrt{n}(max\ \bar{Y}_i - min\ \bar{Y}_i)/\sigma^2$ Let $q_{a,n_{ERR}}$ denote upper α-quantile of the Studentized range dist, aka Tukey's criterion for pairwise comparison at level α
Pairwise comparison procedure	For studentized range dist, Q-statistic for $\mu_i - \mu_j$ is Q_{ij} = $\begin{cases}\frac{\sqrt{n}\lvert\bar{Y}_i - \bar{Y}_j\rvert}{\hat{\sigma}}, if\ n_1 = \cdots = n_a = n \\ \frac{\sqrt{\tilde{n}_{ij}}\lvert\bar{Y}_i - \bar{Y}_j\rvert}{\hat{\sigma}}, otherwise\end{cases}$, where $\tilde{n}_{ij} = \frac{2n_i n_j}{n_i + n_j}$, Contrast $\mu_i - \mu_j$ is significant at level α if Q_{ij} > $q_{a,n_{ERR},\alpha}$
Diff btw Q-statistic & t-statistic	So $\lvert T_{ij}\rvert$ = $Q_{ij}/\sqrt{2}$ t-statistics can be used for pairwise comparison using Tukey's criterion, but $\lvert T_{ij}\rvert$ must be compared w $q_{a,n_{ERR},\alpha}/\sqrt{2}$ i.e. contrast $\mu_i - \mu_j$ is significant at level α if Q_{ij} > $q_{a,n_{ERR},\alpha}$ OR $\lvert T_{ij}\rvert$ > $q_{a,n_{ERR},\alpha}/\sqrt{2}$
Equivalent form in terms of regression coefficient	W conventional defn of dummy vars, the regression coefficients $\beta_i = \mu_i - \mu_1$, i = 2,...,a OR $\mu_i - \mu_j = \beta_i - \beta_j$, i,j > 1 t-statistics: T_{ij} = $\begin{cases}\frac{\hat{\beta}_i}{s.d(\hat{\beta}_i)}, & if\ j = 1 \\ \frac{\hat{\beta}_i - \hat{\beta}_j}{\sqrt{var(\hat{\beta}_i) + var(\hat{\beta}_j) - 2cov(\hat{\beta}_i,\hat{\beta}_j)}}, otherwise\end{cases}$
Tukey's simultaneous CI	For $\beta_i = \mu_i - \mu_1$: $\hat{\beta}_i \pm$ sd($\hat{\beta}_i$)$q_{a,n_{ERR},\alpha}/\sqrt{2}$ For $\beta_i - \beta_j = \mu_i - \mu_j$, i,j > 1: ($\hat{\beta}_i - \hat{\beta}_j$) ± $\sqrt{var(\hat{\beta}_i) + var(\hat{\beta}_j) - 2cov(\hat{\beta}_i, \hat{\beta}_j)}\ q_{a,n_{ERR},\alpha}/\sqrt{2}$
Bonferroni's Mtd	If there are only k contrasts we are interest in, the overall type I error rate α for the k contrasts can be controlled by Bonferroni's Mtd: $\sum_{j=1}^{k}\alpha_j = \alpha$, where α_j is type I error rate for contrast j Each α_j can be specified, but in general just use $\alpha_j = \alpha$/k. Critical value c_α = T_{n-a}(α/k*2). Check $\lvert T_j\rvert$ > c_α for 2-sided test
	Rationale: P($\cup_{j=1}^{k}\{\lvert T_{\boldsymbol{a}}\rvert \geq c_\alpha\}$) ≤ α, where P($\cup_{j=1}^{k}\{\lvert T_{\boldsymbol{a}}\rvert \geq c_\alpha\}$) ≤ $\sum_{j=1}^{k} P(T_{\boldsymbol{a}} \geq c_{\alpha_j})$ ≤ $\sum_{j=1}^{k}\alpha_j$. Thus Bonferroni mtd is a conservative mtd
	p-values: p-value for j[th] test is p_j = kP($T_j \geq T_j^0$) where T_j^0 is observed value of T_j and prob computed under dist of T_j If critical value for all individual test is set at T_j^0, then α_j = P($T_j \geq T_j^0$) for j = 1,...,k and overall type I error rate is α = kα_j i.e. p-value is the overall type I error rate when critical value is observed value of the statistic
Summary	General exploration: Scheffe's criterion Pairwise comparison: all 3 mtds can be used, but Tukey's Mtd more efficient than Scheffe's and Bonferroni's Pre-specified contrasts: Scheffe's and Bonferroni's. Bonferroni more efficient when num of pre-specified contrasts is small For pre-specified contrasts: can just compute both Scheffe's and Bonferroni's, smaller criteria value -> more efficient

LRM w both factor & quant-itative predictor	ANOCOV (analysis of covariance) model is a LRM w both factor & quantitative predictors, $y_i = \beta_0 + \sum_{j=2}^{k}\alpha_j u_{ij} + \gamma x_i + \epsilon_i$ where u_j, j = 2,...,l are dummy variables representing a factor predictor and x is a quantitative predictor, α_j is diff of Y btw type j and type 1 when X is same for both types, γ is effect of X on Y (i.e. when X change by a unit, Y has an expected change of γ) In general ANOCOV model could have more than 1 factor and quantitative predictor. There could be both main and interaction effects
	Traditionally, 1) ANOCOV = comparison of treatment effects of factor predictors, adjusting for effect of certain concomitant variables 2) Comparison of regression fns: r/s btw response var and quantitative predictors is studied in diff categories.
	ANOCOV is based on adjusted SS's (which adjust for effect of concomitant vars). But adjusted SS's have complicated formulae and have similar limitations as traditional ANOVA W regression approach, when estimating factor, effect of concomitant vars is auto adjusted, & avoids drawbacks of traditional ANOVA

Comparison of regression lines	Find whether regression lines have same intercept and whether have same slope -> Use ANOCOV model w interaction btw factor and quantitative predictor: Y = $\beta_0 + \sum_{i=2}^{a} \alpha_i u_i + \beta X + \sum_{i=2}^{a} \xi_i u_i X + \epsilon$, i.e. Category 1: Y = $\beta_0 + \beta X + \epsilon$. Category i ≥ 2: $(\beta_0 + \alpha_i) + (\beta + \xi_i)X + \epsilon$. So just making inference on α_i, and ξ_i
LRM w non-linear predictor terms	In LRM: Y = $\beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$, X need not be diff predictor variables, could be non-linear fns of predictor variables E.g. Y = $\beta_0 + \beta_1 Z + \beta_2 Z^2 + ... + \beta_p Z^p + \epsilon$ | Y = $\beta_0 + \beta_1(1/X) + \epsilon$ (inverse model) | log(Y) = $\beta_0 + \beta_1$log(x) + $\beta_2$log(v)+ ϵ (log model)
Polynomial regression models	$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}^2 + \beta_p x_{2i}^2 + \epsilon_i$, i = 1,...,n (model can have more than 1 predictor variables) $y_i = \beta_0 + \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + \beta_3(x_{1i} - \bar{x}_1)^2 + \beta_p(x_{2i} - \bar{x}_2)^2 + \epsilon_i$ (centralization to ↓ multicollinearity of model)
Piece-wise linear models	R/s btw Y and X might be diff over diff ranges of X Hence use Auxiliary X truncated at point X = c: $\tilde{X} = (X - c)^+ = \begin{cases} X - c, & if\ X - c \geq 0 \\ 0, & otherwise \end{cases}$ Model from $y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ becomes $y_i = \beta_0 + \beta_1 X_i + \alpha_1 \tilde{X}_i + \epsilon_i$ (piece-wise linear in X) Slope of model changes from β_1 to $\beta_1 + \alpha_1$ at point X = c. Can continue adding more auxiliary terms if r/s diff at diff points

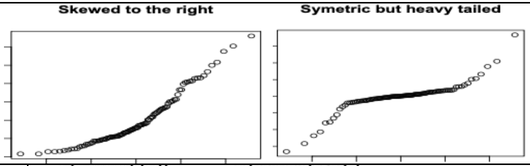
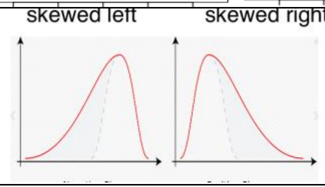
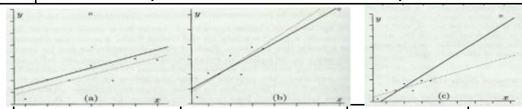
Predictors	Relevant predictor: can explain a proportion of the variation of Y which cannot otherwise be explained by other predictors Relevant predictor is either a causal variable OR a surrogate of certain causal variables not observed Irrelevant predictor is neither a causal variable nor a surrogate of causal variable, BUT might still have correlation w Y. This "fake" correlation aka spurious correlation. Might be that causal variable causes variation in Y and in the irrelevant predictor Irrelevant predictor might have correlation w Y due to data structure (small-n-large-p) where num of predictors > num of observations
	In most cases, model selection is equivalent to variable selection Variable selection is too identify the relevant predictors (causal or surrogate of causal predictors) Variable selection emphasizes accuracy of predictor selection, but model selection emphasizes of accuracy of prediction
Under / Over-Fitting	Under-fitting model: model which does not contain all relevant predictors Over-fitting model: model containing irrelevant predictiors in addition to all relevant predictors
Effect of under/over fitting on LSE	Suppose design matrix is partitioned as X = (X_1, X_2) and includes all relevant predictors. $\mathbf{y}$ = X$\boldsymbol{\beta}$ + $\boldsymbol{\epsilon}$ = $X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$ (full model). $\mathbf{y}$ = $X_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$ (reduced model) Under full model, $\widehat{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T\right)^T = (X^TX)^{-1}X^T\boldsymbol{y}$, E$\widehat{\boldsymbol{\beta}}$ = $\boldsymbol{\beta}$, Var($\widehat{\boldsymbol{\beta}}$) = $\sigma^2(X^TX)^{-1}$. Var($\widehat{\boldsymbol{\beta}}_1$) = $\sigma^2[X_1^TX_1 - X_1^TX_2(X_2^TX_2)^{-1}X_2^TX_1]^{-1}$ Under reduced model, $\widetilde{\boldsymbol{\beta}}_1$ = $(X_1^TX_1)^{-1}X_1^T\boldsymbol{y}$, Var($\widetilde{\boldsymbol{\beta}}_1$) = $\sigma^2(X_1^TX_1)^{-1}$, E$\widetilde{\boldsymbol{\beta}}_1$ = $\boldsymbol{\beta}_1 + (X_1^TX_1)^{-1}X_1^TX_2\boldsymbol{\beta}_2$ Under reduced model, LSE of $\boldsymbol{\beta}_1$ is larger and have a bias of $(X_1^TX_1)^{-1}X_1^TX_2\boldsymbol{\beta}_2$ Under reduced model, var of LSE of $\boldsymbol{\beta}_1$ is smaller since $X_1^TX_1$ > $X_1^TX_1 - X_1^TX_2(X_2^TX_2)^{-1}X_2^TX_1$ (then inverse it to get <) On the other hand, over-fitting increase variance, but decrease bias
Effect of under/over fitting on prediction	Let $\boldsymbol{\beta}_M$ be the parameter of model M w design matrix X_M. The estimator of E$\mathbf{y}$ = $\boldsymbol{\mu}$ under model M is $\widehat{\boldsymbol{y}}_M = X_M\widehat{\boldsymbol{\beta}}_M = X_M(X_M^TX_M)^{-1}X_M^T\boldsymbol{y}$ and E$\widehat{\boldsymbol{y}}_M = X_M(X_M^TX_M)^{-1}X_M^T\boldsymbol{\mu}$ and Var($\widehat{\boldsymbol{y}}_M$) = $\sigma^2X_M(X_M^TX_M)^{-1}X_M^T$ $\sum_{i=1}^{n} Var(\hat{y}_{iM}) = \sigma^2 tr[X_M(X_M^TX_M)^{-1}X_M^T] = M\sigma^2$, where M is num of cols of X_M (usually num of predictors in model M + 1) If we use model M to predict n unobserved y's w the same X_M, the sum of prediction squared error (SPSE) is SPSE = $\sum_{i=1}^{n} E(y_{n+i} - \hat{y}_{iM})^2 = n\sigma^2 + M\sigma^2 + \sum_{i=1}^{n} E(\mu_i - \mu_{iM})^2$ Under fitting increase bias of prediction, decr variance. Over fitting incr variance of prediction, decr bias
Principle of variable selection	Accuracy of prediction is measured by SPSE which consists of a variance and bias component Accuracy of estimate $\hat{\beta}$ is measured by MSE which also consists of a variance and bias component. MSE = $E\left(\hat{\beta} - \beta\right)^2$ = var($\hat{\beta}$) + bias²($\hat{\beta}$) Principle of variable selction is to select variables to balance variance and bias so that SPSE or MSE is minimized SPSE or MSE cannot be practically computed, so model selection criterion uses a surrogate of SPSE or MSE in some sense
Selection Criteria	(see table below)

	R^2 and R_a^2	Mallow's C_p	AIC	BIC	EBIC	CV
Suitable for	models w same num of predictors	Good estimate of σ^2	Minimize Kullback-Leibler dist		models in small-n-large-p problems	based on estimated prediction error
Better	Higher	Lower	Lower	Lower	Lower	Lower

R^2 and R_a^2	$R^2 = \frac{SSR}{SST}$, $R_a^2 = R^2 - \frac{p}{n-p-1}(1 - R^2)$ both estimate proportion of pop variation of response Y explained by predictors in regression model However, R^2 always incr when num of predictors incr. Does not strike a balance btw variance and bias Though R_a^2 not strictly increasing as num of predictors incr, it still incr when predictor having a small contribution to explained variation is added to the model.
Mallow's C_p (Complexity parameter)	Ideally, use $\sum_{i=1}^{n}(y_{n+i} - \hat{y}_{iM})^2$ to estimate SPSE. But don't have y_{n+i} available. So use $\sum_{i=1}^{n}(y_i - \hat{y}_{iM})^2$ to estimate SPSE and find that E$\sum_{i=1}^{n}(y_i - \hat{y}_{iM})^2$ = SPSE - 2|M|σ^2 Adjusting for bias, $\widehat{SPSE} = \sum_{i=1}^{n}(y_i - \hat{y}_{iM})^2 + 2\|M\|\hat{\sigma}^2$ (unbiased estimate of SPSE) Mallow's $C_p = \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(y_i - \hat{y}_{iM})^2 - n + 2\|M\|$. Minimizing C_p is equivalent to minimizing $\widehat{SPSE}$ If model M is correct model, then EC_p ≈ |M|
Akaike's information criterion (AIC)	AIC minimizes the Kullback-Leibler dist btw model M (g_M(y)) and the true model (f(y) approximated by the empirical dist) given by I(f,g_M) = $\int f(y)log\left(\frac{f(y)}{g_M(y)}\right)dy = \int f(y)logf(y)\,dy - \int f(y)log g_M(y)\,dy \approx \int f(y)logf(y)\,dy - \frac{1}{n}\sum_{i=1}^{n} log g_M(y_i)$ AIC approximates the 2nd component above (since can ignore constant = 1st term) AIC = $-2logL\left(\widehat{\boldsymbol{\beta}}_M, \hat{\sigma}_M^2\right) + 2j_M$, where $\widehat{\boldsymbol{\beta}}_M, \hat{\sigma}_M^2$ are MLE under model M, L is likelihood fn and j_M is num of predictors in M For multiple LRM w normality assumption, AIC = $nlog(\hat{\sigma}_M^2) + 2j_M + C$, where $\hat{\sigma}_M^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{iM})^2$ and C = n(ln(2π) + 1)
Bayesian information criterion (BIC)	Let $\mathcal{M}$ be the set of all possible models, Pr a prior probability measure on $\mathcal{M}$. Let Pr(M) be the prior on a model M ∈ $\mathcal{M}$ and $\pi(\boldsymbol{\beta}_M)$ the prior on parameters of model M The marginal density of data $\mathbf{y}$ given model M is m($\mathbf{y}$|M) = $\int L(\boldsymbol{y}, \boldsymbol{\beta}_M)\pi(\boldsymbol{\beta}_M)\,d\boldsymbol{\beta}_M$ The posterior probability of model M is p(M|$\mathbf{y}$) = $\frac{m(y\|M)Pr(M)}{\sum_{M \in \mathcal{M}} Pr(M)m(y\|M)}$ If Pr(M) is taken as constant, BIC = $-2logL\left(\widehat{\boldsymbol{\beta}}_M, \hat{\sigma}_M^2\right) + j_M ln\, n$
Extended Bayesian information criterion (EBIC)	Let model space be partitioned according to num of predictors contained in the models as $\mathcal{M}_0 \cup \mathcal{M}_1 \cup \mathcal{M}_2 \cup$ ..., where $\mathcal{M}_j$ is set of all models containing j predictors. If Pr is the constant prior, then Pr($\mathcal{M}_j$) ∝ $\binom{p}{j}$, where p is total num of predictors When p is large, the prior probability of models w more predictors will be larger than those w less predictors. So BIC will select models w more predictors. EBIC accounts for this problem, by adding a non-constant prior on models EBIC = $-2logL\left(\widehat{\boldsymbol{\beta}}_M, \hat{\sigma}_M^2\right) + j_M ln\, n + 2\gamma ln\binom{p}{j}$, $\gamma = \left(1 - \frac{ln\, n}{2ln\, p}\right)$

Cross Validation (CV)	Model M fitted using data ($\mathbf{y}$, X). To assess goodness of model, ideally use another data set, $(\tilde{\mathbf{y}}, \tilde{X})$ and validate by prediction error $\lVert\tilde{\mathbf{y}} - \tilde{X}\hat{\boldsymbol{\beta}}\rVert^2$. Practically, same dataset is split into training and testing data. So wastes information and is against principle of sufficiency Leave-out-one CV: Training data is n-1 observations, test data is 1 observation k-fold CV: Whole data divided into k parts, each time, one part is test data, remaining k-1 parts for training data Leave-out-one CV score: CV = $\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}^{(-i)}\right)^2$, where $\hat{\boldsymbol{\beta}}^{(-i)}$ is the estimate of $\boldsymbol{\beta}$ by leaving out the i^{th} data point (y_i, $\mathbf{x}_i$) k-fold CV score: $CV_k = \frac{1}{k}\sum_{i=1}^{k}\left(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}^{(-i)}\right)^2$, where ($\mathbf{y}_j$, X_j) is j^{th} part of data (testing data) & $\hat{\boldsymbol{\beta}}^{(-i)}$ = estimate obtained by training data
Model selection strategy	Naive method: all subset selection. Not practical as for p predictors, there are 2^p possible models
	1) Remove redundant predictors (when p is not very large) i. Fit full model, remove all predictors w p-value bigger than a certain level α ii. Fit full model, remove predictor w largest p-value which is > α, then re-fit model w remaining predictors, repeat this until no predictor has p-value > α If covariates are not highly correlated, the 2 options produce the same selected model. If high correlation among covariates exists, ii. preferred
	2) Forward selection (sequential procedure) Starts w null model M_0 w no predictors, then add predictors 1 at a time, choosing predictor having largest contribution to reduce the residual sum of squares Compare new model w old model by certain criterion. If new model better than old model -> continue; otherwise stop. Criterion can be any except R^2 and R_a^2
	3) Backward Selection (inverted sequential method) Start w full model M_F w all predictors, then reduce model by removing predictors one at a time, choosing predictor w smallest contribution to reduce residual sum of squares to be removed. Compare reduced model w previous model by AIC. If AIC of new < AIC of old -> continue; otherwise stop
	4) Stepwise selection (mixture of forward & backward selection. Can be done upwards OR downwards) Upward stepwise selection; Start w null model. Add predictor to model, perform backward procedure until no predictor can be removed. Proceed to next forward step. Repeat Downward stepwise selection: Start w full model. Remove predictor from model, perform forward procedure until no predictors can be added. Proceed to next backward step. Repeat
Penalized likelihood approach	For LRM $\mathbf{y} = X\boldsymbol{\beta} + \epsilon$, the penalized likelihood approach select variables by minimizing $\frac{1}{n}\lVert\mathbf{y} - X\boldsymbol{\beta}\rVert_2^2 + p_\lambda(\boldsymbol{\beta})$, where p_λ is the penalty function, λ is the penalty parameter whose value is to be chosen Procedure: specify sequence of λ values, at each value, carry out the penalized minimization, which yields a model w certain selected variables. Selection criteria is used to select the model If purpose to obtain model for prediction -> CV. If purpose to identify important variables -> EBIC
Common penalty functions	1) LASSO penalty: $p_\lambda(\boldsymbol{\beta}) = \lambda\sum_{j=1}^{p}\lvert\beta_j\rvert$ 2) Adaptive LASSO penalty: $\lambda\sum_{j=1}^{p} w_j\lvert\beta_j\rvert$, where w_j is taken as $\lvert\hat{\beta}_j\rvert^{-1}$. If $p \ll n$, $\hat{\beta}_j$ being the OLS (ordinary least square) estimator in multiple LRM. If p is close to or > n, $\hat{\beta}_j$ is the OLS estimate is the marginal LRM SCAD penalty: $p_\lambda(\lvert\beta\rvert) = \lvert\beta\rvert$ for $\lvert\beta\rvert$ near 0, and equals a constant C for large $\lvert\beta\rvert$, the two parts are connected by a smooth function MCP penalty: for large $\lvert\beta\rvert$, it is a constant. Smoothly decreases to 0 w $p_\lambda(\beta) = \lvert\beta\rvert$ as its asymptote when $\lvert\beta\rvert$ approaches 0
Rationale of penalized likelihood approach	E.g. LASSO (least absolute shrinkage and selection operator) estimates $\boldsymbol{\beta}$ by minimizing $\frac{1}{n}\lVert\mathbf{y} - X\boldsymbol{\beta}\rVert_2^2 + \lambda\sum_{j=1}^{p}\lvert\beta_j\rvert$ If λ = 0, LASSO estimator is same as LSE. If $\lambda = \infty$, all components of $\boldsymbol{\beta}$ are estimated as 0 For a certain nonzero λ, some of the components will be estimated as nonzero, and others 0. The nonzero ones are shrunk version of LSE Variables w nonzero estimated coefficients are the selected variables

Model diagnostics	Assumptions made for model might not be true, leading to discrepancies. There are 2 types of discrepancies – systematic and local	
	Fitted values are $\hat{y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$, i = 1,...,n. Hat matrix is $H = X(X^TX)^{-1}X^T$ Hat values (i^{th} diagonal elem of H), $h_{ii} = \boldsymbol{x}_i^T(X^TX)^{-1}\boldsymbol{x}_i$ = hat value of i^{th} observation, where $\boldsymbol{x}_i^T$ is the i^{th} row of X	
	Partitioning X as X = ($\mathbf{1}$ Z), $(X^TX)^{-1} = \begin{pmatrix} n & \mathbf{1}^TZ \\ Z^T\mathbf{1} & Z^TZ \end{pmatrix}^{-1} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ $A_{22} = \left[Z^T\left(I - \frac{\mathbf{1}\mathbf{1}^T}{n}\right)Z\right]^{-1}$ = sample covariance matrix, $A_{11} = \frac{1}{n} + \frac{\mathbf{1}^TZ}{n}A_{22}\frac{Z^T\mathbf{1}}{n}$, $A_{12} = -\frac{1}{n}\mathbf{1}^TZA_{22}$, $A_{21} = A_{12}^T$ Let i^{th} row vector $\mathbf{x}_i$ of X be $(1, z_i)^T$ and $\bar{\mathbf{z}} = \frac{Z^T\mathbf{1}}{n}$. Then $A_{11} = \frac{1}{n} + \bar{\mathbf{z}}^TA_{22}\bar{\mathbf{z}}$ Then $h_{ii} = \boldsymbol{x}_i^T(X^TX)^{-1}\boldsymbol{x}_i = \frac{1}{n} + (\mathbf{z}_i - \bar{\mathbf{z}})^TA_{22}(\mathbf{z}_i - \bar{\mathbf{z}})\bar{\mathbf{z}} = \frac{1}{n}$ + Mahalanobis dist of i^{th} observation in x-space to centroid of sample	
	Pearson's residuals are $e_i = y_i - \hat{y}_i$, i = 1,...,n. Let $\mathbf{e} = (e_1,...,e_n)^T$. Then $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (I - H)\mathbf{y}$. $E\mathbf{e} = \mathbf{0}$, var($\mathbf{e}$) = $\sigma^2(I - H)$. var(e_i) = $\sigma^2(1 - h_{ii})$ Under normality assumption, $\mathbf{e} \sim N(\mathbf{0}, \sigma^2(I - H))$	
	Studentized residuals, $r_i' = \frac{e_i}{sd(e_i)} = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$ Studentized deleted residuals, $r_i^* = \frac{y_i - \hat{y}_{i(i)}}{sd(y_i - \hat{y}_{i(i)})} = \frac{y_i - \hat{y}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$, where $\hat{y}_{i(i)}$ is the predicted value by fitted model w i^{th} observation deleted, $\hat{\sigma}_{(i)}$ is the counter part of $\hat{\sigma}$ when i^{th} observation is deleted Cook's distance, $d_i = \left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)^T(X^TX)\left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)/(p\hat{\sigma}^2)$, where $\hat{\boldsymbol{\beta}}_{(i)}$ is estimate of $\boldsymbol{\beta}$ when i^{th} observation is removed from data Variance Inflation Factor (VIF). Let R_k^2 be the coefficient of determination of model $X_k = \beta_0 + \sum_{j\neq k}\beta_j\boldsymbol{X}_j + \epsilon$. VIF of X_k is $VIF_k = \frac{1}{1-R_k^2}$	
Systematic discrepancies	Caused by regression fn not linear, error terms don't have constant variance, error terms not indep, error terms don't have normal dist, impt predictors are omitted from model. Use residual plots to check for systematic error	
	If no discrepancies, residuals would appear like iid random errors w mean 0. In any residual plot, points are scattered evenly within a horizontal band around 0	
Check non-linearity	Plot Pearson's residual against fitted values Plot Pearson's residual against predictor variables Scatter plot of response against predictor variables	If any of the plots show a non-linear trend -> regression fn is not linear For residual vs predictors: no trend -> no obvious discrepancy in regression fn
Check homogeneity	Plot Pearson's residual against fitted values Plot Pearson's residual against predictor variables	If vertical range of residuals have obvious change along x-axis -> variances are not constant / not homogeneous

Check independence	Plot residual against time/space	If indep -> should be constant horizontal trend
Check normality	Plot of studentized residuals through dist plot (box plot, histogram) OR normal probability plot of residuals OR QQ plot If normality holds, points in QQ plot shld fall on straight line y = x	
Heavy tailed pattern	If dist of r.v. Y is skewed to the right (positive skew) relative to normal dist, then P(Y ≤ c) ≤ P(Z ≤ c) for all c Let y_q and z_q denote q-quantile of Y and Z, then P(Y ≤ y_q) = P(Z ≤ z_q) ≥ P(Y ≤ z_q). Then P(Y ≤ y_q) ≥ P(Y ≤ z_q) Hence y_q ≥ z_q for all q. In Q-Q plot where y_q is plotted against z_q, the point (z_q, y_q) is above the point (z_q, z_q)	
Check missing predictors	Plot residual against other predictors not included in the model	If 1 of the plot show a trend -> that predictor is missing
Outliers	Leverage: whether point is far away from major cluster in x-space Since $\sum_i h_{ii}$ = Tr(H) = p. Point is high leverage if $h_{ii} > \frac{2p}{n}$ Consistency: whether point is consistent in terms of fitting in the (x,y)-space Studentized deletion residuals are the standardized prediction errors, Find points w highest $\lvert r_i^* \rvert$ values -> possible outliers Influence: whether point highly affects fitting of model Find points w highest Cook's distance -> possible outliers	
Assessment of outliers	Informal test of outliers: normal probability plots of studentized deletion residual, leverage h_i and Cook's distance d_i Formal test: To assess (y_i, $\mathbf{x}_i$), introduce the dummy variable u = $\begin{cases} 1, for\ i^{th}\ unit \\ 0, otherwise \end{cases}$. Significance of coefficient of u in the linear predictor indicates i^{th} point is an outlier.	

	a	b	c
Leverage	low	high	high
Consistency	No	Yes	No
Influence	low	low	high

1. LRM w unequal variances	When ϵ_i's don't have common variance, an unequal variance model is considered: $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ has the variance matrix $\Sigma = \sigma^2 \begin{pmatrix} w_1^{-1} & 0 & \cdots & 0 \\ 0 & w_2^{-1} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & w_n^{-1} \end{pmatrix}$, where w_i's are unequal weights (if common variance: $\sigma^2 I$)
	If w_i's are known, the unequal variance model can be transformed into an equal variance model. Let W = $\begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & w_n \end{pmatrix}$. Multiply $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ by $W^{1/2}$, then $W^{1/2}\boldsymbol{y} = W^{1/2}X\boldsymbol{\beta} + W^{1/2}\boldsymbol{\epsilon}$. Let $\tilde{\boldsymbol{y}} = W^{1/2}\boldsymbol{y}$, $\tilde{X} = W^{1/2}X$, $\tilde{\boldsymbol{\epsilon}} = W^{1/2}\boldsymbol{\epsilon}$ Then var($\tilde{\boldsymbol{\epsilon}}$) $= W^{1/2}\Sigma W^{1/2} = \sigma^2 I$. And model $\tilde{\boldsymbol{y}} = \tilde{X}\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}$ is a constant variance model
	Minimizing $\lVert \tilde{\boldsymbol{y}} - \tilde{X}\boldsymbol{\beta} \rVert_2^2$, we obtain the estimate of $\boldsymbol{\beta}$ as $\widehat{\boldsymbol{\beta}}_W = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{\boldsymbol{y}} = (X^TWX)^{-1}X^TW\tilde{\boldsymbol{y}}$ = weighted LSE (WLSE) $\lVert \tilde{\boldsymbol{y}} - \tilde{X}\boldsymbol{\beta} \rVert_2^2$ can be expressed explicitly as $\sum_{i=1}^n w_i(y_i - \boldsymbol{x}_i^T\beta)^2$. The weight w_i reflect the relative importance of $(y_i - \boldsymbol{x}_i^T\beta)^2$ in the estimation. The larger the variance of i^{th} term, the smaller the corresponding weight, since weight is inversly proportional to the variance
	For WLSE $\widehat{\boldsymbol{\beta}}_W$, E$\widehat{\boldsymbol{\beta}}_W$ = $\boldsymbol{\beta}$, var($\widehat{\boldsymbol{\beta}}_W$) = $\sigma^2(X^TWX)^{-1}$. And $\widehat{\boldsymbol{\beta}}_W \sim N(\boldsymbol{\beta}, \sigma^2(X^TWX)^{-1})$ σ^2 is estimated as $\hat{\sigma}^2 = \frac{1}{n-p-1}\lVert \tilde{\boldsymbol{y}} - \tilde{X}\widehat{\boldsymbol{\beta}}_W \rVert_2^2 = \frac{1}{n-p-1}\sum_{i=1}^n w_i(y_i - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_W)^2$. Inference on $\boldsymbol{\beta}$ is made in same way as in normal LRM
Estimation of unknown weights	If no replicates of predictor values, Weights can be estimated by the following procedure: 1. Fit regression model by unweighted least squares and obtain residuals **r** and fitted values $\hat{\boldsymbol{y}}$ 2. Regress log of absolute residual on log of fitted values, e.g. ln $\lvert r_i \rvert$ = $\alpha_0 + \alpha_1 ln\ \hat{y}_i + e_i$. If smallest $\hat{y}_i$ < 0, replace $\hat{y}_i$ by $\hat{y}_i - min\ \hat{y}_i + c$, for some positive constant c 3. Weights are estimated as $w_i = \hat{y}_i^{-2\hat{\alpha}_1}$ OR $e^{-2\hat{\alpha}_0}\hat{y}_i^{-2\hat{\alpha}_1}$. Constant $e^{-2\hat{\alpha}_0}$ don't really matter
	If there are replicates of predictor values, sample variance can be used in estimation of weights 0. Naive method: weight = 1/s^2, s = sample SD 1. Estimate variance as a function of the mean by using the regression model: ln s_i = $\alpha_0 + \alpha_1 ln\ \lvert\bar{y}_i\rvert + e_i$, where s_i is sample sd, and $\bar{y}_i$ is sample mean 2. The weight for the i^{th} predictor value is $w_i = \lvert\bar{y}_i\rvert^{-2\hat{\alpha}_1}$
2. Multi-collinearity & its effects	Correlation among predictor variables such as pariwise correlation, linear dependence of one predictor on another predictor,... If 1 predictor is perfectly linearly dependent on the other predictors, X^TX would be singular, i.e. non-invertible Practically, although perfect linear dependence will not occur, high multicollinearity can cause X^TX to be nearly singular (i.e. having large condition number $\lambda_{max}/\lambda_{min}$), which renders the LSE extremely unstable Serious multicollinearity greatly increases variance of LSE and make LSE inaccurate and useless
Informal Diagnostics for multi-collinearity	Large changes in estimated regression coefficients when a predictor is added/deleted, or an observation is altered or deleted Nonsignificant results in individual test on the regression coefficients for known important predictor variables Estimated regression coefficients w an algebraic sign that is opp of that expected from theoretical consideration or prior experience Large coefficients of sample correlation btw pairs of predictors in the correlation matrix (linear trend, correlation > 0.6)
1) Formal diagnostic – VIF	X_{-j} being the submatrix of X w/o its j^{th} column and R_j^2 is the coefficient of multiple determination of $\mathbf{x}_j$ is regressed on X_{-j} If $\mathbf{x}_j$ is uncorrelated w X_{-j} -> R_j^2 = 0 and the variance of $\hat{\beta}_j$ = σ^2/SST_j If $\mathbf{x}_j$ is correlated w X_{-j} -> the variance is inflated by a factor VIF_j = $(1 - R_j^2)^{-1}$ (Variance inflation factor)
2) Ridge regression	Since consequence of multicollinearity is X^TX nearly singular, to remedy: add diagonal matrix to X^TX, i.e. $X^TX + \lambda I$ (which is invertible) Ridge regression estimator $\widehat{\boldsymbol{\beta}}_{RIG} = (X^TX + \lambda I)^{-1}X^T\boldsymbol{y}$, where λ > 0 is a parameter to be chosen $\widehat{\boldsymbol{\beta}}_{RIG}$ is also the minimizer of the penalized sum of squares: $\sum_{i=1}^n(y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^p \beta_j^2 = \lVert \boldsymbol{y} - X\boldsymbol{\beta} \rVert^2 + \lambda\lVert\boldsymbol{\beta}\rVert^2$
	E$\widehat{\boldsymbol{\beta}}_{RIG}$ = $(X^TX + \lambda I)^{-1}X^TX\boldsymbol{\beta} = \beta - \lambda(X^TX + \lambda I)^{-1}\beta$ -> estimator is biased. Bias is $\beta - \text{E}\widehat{\boldsymbol{\beta}}_{RIG} = \lambda(X^TX + \lambda I)^{-1}\beta$

Let Q be the orthogonal matrix s.t. $X^TX = Q\Delta Q^T$, where Δ = Diag($\tau_0, \tau_1, \ldots, \tau_p$). Thus $\beta - \mathrm{E}\widehat{\boldsymbol{\beta}}_{RIG} = Q \begin{pmatrix} \frac{\lambda}{\tau_0+\lambda} & 0 & \cdots & 0 \\ 0 & \frac{\lambda}{\tau_1+\lambda} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \frac{\lambda}{\tau_p+\lambda} \end{pmatrix} Q^T\boldsymbol{\beta}$

Let $\tilde{\beta}_j$ denote the j[th] component of $Q^T\boldsymbol{\beta}$. We have $\left\|\beta - \mathrm{E}\widehat{\boldsymbol{\beta}}_{RIG}\right\|^2 = \sum_{j=0}^{p}\left(\frac{\lambda}{\tau_j+\lambda}\right)^2 \tilde{\beta}_j$. Thus bias increase as λ increase

Thus tr(var($\widehat{\boldsymbol{\beta}}_{RIG}$)) = $\sigma^2 \sum_{j=0}^{p} \frac{\tau_j}{(\tau_j+\lambda)^2}$. The variance decrease as λ increase. All other cov = 0

Sum of mean squares errors of $\boldsymbol{\beta}$ is given by MSE = $\sum_{j=0}^{p} MSE(\widehat{\boldsymbol{\beta}}_j) = \left\|\beta - \mathrm{E}\widehat{\boldsymbol{\beta}}_{RIG}\right\|^2 + TR(\mathrm{Var}(\widehat{\boldsymbol{\beta}}_{RIG}))$.
Need to get balance btw bias and variance by minimizing MSE
MSE cannot be readily used as a criterion as it involves unknowns $\boldsymbol{\beta}$ and σ^2. Can select λ by Cross validation (CV)
Let $\widehat{\boldsymbol{\beta}}_{-i}(\lambda)$ be the ridge regression estimate of $\boldsymbol{\beta}$ w parameter λ by deleting the i[th] observations. The CV score is given by CV(λ) = $\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_{(i)}^T\boldsymbol{\beta}_{-i}(\lambda)\right)^2$, where $\boldsymbol{x}_{(i)}^T$ is the i[th] row vector of the design matrix X
Best λ is the minimizer of CV(λ)

Note ridge regression mainly used for building model for prediction. Cannot be used to assess importance or effects of predictor variables.
The estimates from model cannot be used to construct CI or conduct hypo testing
If need to make inference on effects of predictors -> use strategy of removing predictors w large VIF

3. Non-normality (remedy w variable transformation)

If normal dist -> variance don't depend on mean (i.e. constant variance in regression models)
The violation of normality usually goes tgt w violation of constancy of variance
A variance stabilization transformation can help to rectify both discrepancy in normality and constancy of variance
If r/s btw variance and mean is known -> the desired variance stabilization transformation can be derived
To find the transformation -> use Box-Cox transformation

If variance σ^2 depend on mean μ, i.e. σ^2 = V(μ), a transformation can be found s.t. variance of transformed variable is ≈ indep of mean
Let h(Y) be a transformation. Use taylor series to expand h(Y) at μ, as h(Y) ≈ h(μ) + h'(μ)(Y-μ), where h(μ) is a constant
Treating h(μ) as mean of h(Y), var(h(Y)) = E[h(Y) - h(μ)]^2 ≈ [h'(μ)]^{2}E(Y - μ)^2 = [h'(μ)]^{2}V(μ)
WLOG, setting [h'(μ)]^{2}V(μ) = 1, h'(μ) = $\frac{1}{\sqrt{V(\mu)}}$ -> h(μ) = $\int \frac{1}{\sqrt{V(\mu)}} d\mu$, aka Variance Stabilization transformation

Box-Cox transformation

If Variable transformation don't work?
For certain non-normal cts r.v., transformation cannot be determined solely by data type
Box-Cox transformation: h(Y) = $\frac{Y^\lambda - 1}{\lambda}$. λ can be determined by data
When λ = 0, the Box-Cox transformation is given by h(Y) = ln(Y), since $\lim_{\lambda\to 0}\frac{Y^\lambda-1}{\lambda}$ = ln(Y)

Let $\sigma_i \propto \mu_i^\alpha$, where σ_i and μ_i are the SD and mean of i[th] treatment effect. The λ in Box-Cox transformation is determined by variance stabilization transformation

$\sigma_i \propto \mu_i^\alpha$	α (just use closest half int, e.g. α = 0.04 -> use 0.05)	λ = 1 - α	Transformation
$\sigma_i \propto \mu_i^3$	3	-2	reciprocal squared
$\sigma_i \propto \mu_i^2$	2	-1	reciprocal
$\sigma_i \propto \mu_i^{3/2}$	3/2	-1/2	reciprocal sqrt
$\sigma_i \propto \mu_i$	1	0	log
$\sigma_i \propto \mu_i^{1/2}$	1/2	1/2	sqrt
$\sigma_i \propto$ constant	0	1	no transformation
$\sigma_i \propto \mu_i^{-1/2}$	-1/2	3/2	3/2 power
$\sigma_i \propto \mu_i^{-1}$	-1	2	square

Determination of α
Method 1: If observations are grouped, for each group, compute s_i and $\bar{y}_i$. Fit regression model ln $s_i = \beta_0 + \beta_1$ln $\bar{y}_i + e_i$
If observations not group, fit ln $|r_i| = \beta_0 + \beta_1$ln $\hat{y}_i + e_i$
For both, α = estimate of β_1
Method 2: Only for grouped observations. Select a few α values, say α_k, k = 1,...,K. For each k, compute $R_k = \max_i \frac{s_i}{\bar{y}_i^{\alpha_k}} / \min_i \frac{s_i}{\bar{y}_i^{\alpha_k}}$
Select α_k with smallest R_k

Direct determination of λ
For grouped observations: select a few λ values, say λ_k, k = 1,...,K
For each k, make the transformation $y_{ij} \to y_{ij}^{\lambda_k}$. With the transformed data, compute $s_{\lambda_k i}^2$, i = 1,...,g
Select λ_k s.t $\max_i s_{\lambda_k i}^2 / \min_i s_{\lambda_k i}^2$ is closest to 1
For non-grouped observations: select a few λ values, say λ_k, k = 1,...,K
For each k, make the transformation $y_{ij} \to y_{ij}^{\lambda_k}$. Analyze the regression models w $y_i^{\lambda_k}$ as response variable.
Select λ_k s.t MSE(λ_k) is smallest