|  |  |  |  |
| --- | --- | --- | --- |
| Special Matrix | | Symmetric matrix A is +ve/(-ve) definite if for any vector x, xTAx > (<)0. ≥ (non -ve); ≤ (non +ve) definite  If A is a sq matrix and ATA = I, then A is orthogonal matrix -> rows/cols of A form orthonormal basis for Rn -> 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, = (A), where denotes largest eigenvalue  If A is non-negative definite and B is non-singular (i.e. has inverse, det(B) = 0), = (AB-1) | |
| Descriptive quantities | | | Var of X is = E(X - E(X))2 = . Sample var is  Let Y be another r.v. w observed sample y1,..., yn. 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  Suppose X has a cts dist. The (1 - )-quantile (upper -quantile), is defined s.t. F() = 1 - or 1 - F() = |
| Mean: E = . Var: Var(X) = E(X2) - [E(X)]2  Var = in general; Var = if Xj's are un-correlated  Covariance: Cov(X, Y) = E(XY) - E(X)E(Y). Cov = |
| Sample var: - . Sample cov: = - |
| 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(X1), ..., E(Xp))T  Var(AX + **b**) = AVar(X)AT where Var(X) = Cov(Xi, Xj) = E(X - E(X))(X - E(X))T. Note Cov(Xi, Xi) = Var(Xi)  Cov(AX + **b**, BX + **c**) = AVar(X)BT. Cov(AX + **b**, BY + **c**) = ACov(X, Y)BT where Cov(X, Y) = Cov(Xi, Yj) = 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|, ) = where is the mean and is the var of the dist  Normal dist is denoted by N(, ). If = 0, = 1, it is the standard normal dist. Normal dist is symmetric about its mean  If X ~ N(, ), then Z = (X - )/ ~ N(0,1) := standardisation |
| Uni-variate Normal dist | | Let **Y** = (Y1, ..., Ym)T be random vector whose components are iid standard normal variables. Let A be q x m constant matrix and a constant vector. Dist of **X** = A**Y** + is a q-dimensional multivariable normal dist w mean E(**X**) = and var matrix ∑ = AAT, and denoted by N(, ∑)  ∑ = . Normal dist is uniquely determined by its mean and var matrix ∑  pdf of N(, ∑) is f(**x**|, ∑) =  If X is a multivariate normal vector, then for any constant matrix B, B**X** has a multivariate normal dist N(B, B∑BT)  constant vector **c**, the LC **c**T**X** has a univariate normal dist N(**c**T, **c**T∑**c**) and any component of **X** is an univariate normal var  If **X**~N(, ∑), then **Z** = ~N(**0**, I), i.e. components of **Z** are iid N(0, 1) variables | |
| Multi-variate normal dist | Let **Z** = (Z1,..., Zm)T. Suppose Zj's are iid N(0,1) variables. Dist of **Z**T**Z** = is called the -dist w d.f. m and denoted by  Suppose Z~N(0,1), U~, Z and U are indep. The dist of Z/() is called the t-dist w d.f. m and denoted by tm  Suppose U~, V~, U and V are indep. Dist of is called the F-dist w d.f. m and n and denoted by Fm,n  If **X**~N(, ∑), then W = (**X** - )T~  If Z~N(0,I), A is symmetric and idempotent, then **Z**TA**Z**~, where r = rk(A) = tr(A) | | |
| , t and F - dist | | Let be param of interest. Suppose = g(m1, m2,...) (1st moment, 2nd moment,...). Then MME of is obtained by replacing the theoretical moments in the fn w the corresponding sample moments, i.e. = g(, ,...)  For any dist, the var = m2 - m12 (E(X2) - [E(X)]2), its MME is given by = - . MME is not unique. | |
| Mtd of Moment estimation (MME) | | | If X has pdf f(x, ), given the observation x1,...,xn of a random sample, the log likelihood fn of is defined as () = . The MLE of is value of that maximizes the log-likehood fn.  E.g. let (x1,...,xn) be observation of random sample from N(, ).  Likelihood fn: ( is prediction here, i.e. = yi - )  The log likelihood fn of (, ) is then (, ) = = – log(2π) –  MLE of and are obtained by maximizing (, ) and are given by = , = |
| Maximum likelihood estimation (MLE) | | | Null hypothesis H0 and alternative hypothesis H1. The 2 hypotheses are mutually exclusive  A test statistic, T(x) is used. If T(x) ≥ c for a predetermined constant c -> reject H0. If not -> don't reject H0  To control the type I error rate (reject H0 when H0 true) at a given level , i.e. choose c s.t. P(T(**X**) ≥ c|H0) ≤ -> reject H0  H0 should be s.t. type I error is more serious. If still not clear, H0 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 ≤ ≤ 1  Pearson's correlation, the theoretical correlation is = , the sample correlation is = = corr(X, Y)  If > 0: move in same dir. If < 0: move in opp dir. Both DO NOT imply a causal r/s  If = 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 = + X + ,  E(Y) = + X is called the regression function (E(Y|X) more technically correct way to write it)  W observations (xi, yi), observed simple LRM is yi = + xi + , i = 1,...,n. Error term is diff for ea observation | | | |
| Assumptions of simple linear regression model (LRM):   |  |  | | --- | --- | | 1. xi and are indep. Note indep un-correlated, not necessary. (Unless normal var, then indep <-> un-correlated) | | | 2. have mean 0 | 3. are pairwise un-correlated, i.e. Cov(, ) = 0 | | 4. have common variance (Homogeneity) | 5. have a Normal dist (Normality) |   LRM is then called a Normal LRM. Assumptions can be simply stated, , iid ~N(0, ) in addition to (1) | | | |
| From Y = + X + , 1) E(Y) = + E(X) 2) Cov(X, Y) = Var(X)  So = = and = - (exact values of , )  (, ) will then be the soln to (i.e. error term minimised)  So regression fn + X is best linear approximation of X to Y | | | |
| Estima-tion of LRM | LSE of Property (, ) minimizes gives rise to the least square estimation (LSE), which estimate the params by minimizing the sum of squares, Q = = SSE  To find LSE of , , = -2, and = -2 and equate both to 0  Summary: = , = , where = and = | | | |
| LSE of = E(). =  SSE (sum of square of error) = = and estimated by = s2 = SSE/(n-2)  Estimator of is is called the residual standard error | | | |
| Fitted/estimated regression fn: = + X  Fitted values: = , i = 1,...,n | | | Residuals: ei = yi - , i = 1,...,n  Value = + 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 yi = + | | | |
| ANOVA table for SRM   |  |  |  |  |  | | --- | --- | --- | --- | --- | | source | df (deg of freedom) | SS (sum of sq) | MS (mean of sq) | F (f ratio) | | Regression | 1 | SSR = | MSR = SSR/1 | MSR/MSE | | Error | n-2 | SSE = | MSE = SSE/(n-2) |  | | Total | n-1 | SST = |  |  | | | | |
| 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 = R2 - where p is num of predictors and equals 1 for simple LRM  R2 is strictly increasing as p increases. But does not necessarily incr as p incr | | | |
| = , = , SSR = , R2 = corr (Y, )2 = corr(Y, X)2 = = = =  Summary results explained:   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Estimate | Std. Error | t value | Pr(>|t|) | | (Intercept) |  | s() | = | p-value for 2-sided test on | | X |  | s() | = | p-value for 2-sided test on | | Residual standard error: | | | = | | | Multiple R-squared: | | R2 = = 1 – = | Adjusted R-squared: |  | | F-statistic: | | MSR/MSE = ()2 | p-value | p-value of the significant F test | | | | |
| = + X. Using R2, can explain how much of variation of Y is due to X  is estimated amt of change in expectation of Y when X incr by an unit amt  Can predict Y\* w new observation X\* using + X\*  Fitted regression function: = + X. We can say Y incr/decr as X incr. When X incr by 1 unit, Y incr/decr by  X explaines about (R2\*100)% of the variation in Y | | | |
| Theoretical Properties of LSE | | | Unbiasedness of s2 / as an estimator of : Thus Es2 = | |
| Properties of estimated/fitted regression fn:  The estimated regression fn = + X is an unbiased estimator of EY = + X, i.e. E() = E(Y)  var() = . OR using = + X, var() = var() + 2XCov() + X2var()  For the prediction at a new value X, the prediction mean square error is: E(Y-)2 = var() + var() = | |
| Dist of LS estimators: Normality condition is assumed. , are LC of Y1, ..., Yn  By property of normal dist, , are normally distributed: ~N, ~N  (n-2) (where s2 = ) ~-dist w df n-2 & is indep from , (verification given later) | |
| Significant test: test whether or not there is a linear regression r/s btw the response var & the covariate  Null hypothesis, H0: no such r/s. Alternative hypothesis, H1: r/s exist  test statistic is F-statistic, where F = (intuitively, if variation caused by covariate > variation caused by error r/s exist)  Under H0, F~F-dist with df 1 and n-2, since SSR and SSE indep and each follows a -dist  If F > upper quantile f1, n-2(), H0 is rejected at level , otherwise not rejected OR P(F > f1, n-2()) < = P(type 1 error) | |
| t-statistic for the inference on  Dist of cannot be used directly for inference, since its var involves which is unknown  Let s2() = . Define = = t-dist ~ tn-2. | |
| Statistical Inference for simple LRM | | | Two-sided test for : H0: = 0 and H1: ≠ 0  Test-statistic: = ~ tn-2 under H0  Significance level (usually taken as 0.05 or 0.01). If || > tn-2(/2), then reject H0. Otherwise, don't reject H0  OR p-value = p = 2P(|tn-2| > ||) (T is observed value here, above was r.v). If p < , reject H0; otherwise, don't reject H0  For simple LRM, the two-sided test for = p-value of F-statistic. F-statistic here = (t-statistic)2 | |
| One-sided test for  1) H0: ≤ 0, H1: > 0. (Equal sign w H0). If > tn-2(), OR p = P(tn-2 > ): reject H0; otherwise don't  2) H0: ≥ 0, H1: < 0. If < tn-2(), OR p = P(tn-2 < ): reject H0; 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 = | |
| From dist of , 100(1 - )% CI for is [ - tn-2(/2)s(), + tn-2(/2)s()]. CI are used for 2-sided tests | |
| 100(1-)% lower confidence bound: ≥ - tn-2()s(). This corresponds to 1-sided test (H0: ≤ 0, H1: > 0)  100(1-)% upper confidence bound: ≤ + tn-2()s(). This corresponds to 1-sided test (H0: ≥ 0, H1: < 0) | |
| The inference on  = , where s2() =  Dist of also tn-2. Inference is similar to . E.g. 95% CI for is ± tn-2(/2)s() | |
| Prediction  CI of E(Y) when predictor value is xh: , where where is the predicted value for xh  Prediction interval of Ynew when predictor value is xh: , where var(Ynew) =  var(Ynew) = var() + , where = + xh  Manual computation: var() = var() + var() + 2xhcov(, ). values and can be found from summary results | |
|  | | LSE as Method of moments estimation (MME): just replace theoretical moments w sample moments  From = , = - LSE is then = and = - , 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  = MLE of is given by = = s2  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: = a1x1 + ... + anxn = **a**T**x**, where **a** = (a1,...,an)T, **x** = (x1,...,xn)T  Matrix form of quadratic fns: = **x**TA**x**, where A =  Matrix form of differentiation: Let f(**x**) be a multivariate fn. Define = ,  then = **a**, and = 2A**x** if A is symmetric, (A + AT)**x** otherwise | | | | |
| In multiple LRM, find r/s btw response var Y and p predictor variables/covariates **X** = (X1,..., Xp)  Then multiple LRM is Y = + X1 + ... + Xp + , where is random error w mean 0  Then EY = + X1 + ... + Xp is the multiple linear regression fn. (Or more strictly speaking EY is E(Y|X)  For n observations, yi = + + ... + + , i = 1,...,n | | | | |
| Let X = , , **y** = ,  **=** . Then = OR **y** = X+ . X aka design matrix | | | | |
| |  |  | | --- | --- | | Assumptions for multiple LRM (similar to simple LRM) | 1. X1,...,Xp and are indep. | | 2. have mean 0 | 3. are pairwise un-correlated, i.e. Cov(, ) = 0 | | 4. have common variance (Homogeneity) | 5. have a Normal dist (Normality) |   Let denote mean of Y, denote mean vector of X = (X1,...,Xp), : covariance vector btw Y and X, : covariance matrix of X and = (,..., ). By assumption1, = , = - | | | | |
| Least sq estimate for multiple LRM | | | Estimate , , ..., by minimizing Q = = (norm)  LSE of is = (XTX)-1XT**y**  = = = | | | |
| Hat matrix & properties | | H = X(XTX)-1XT is the hat matrix of X. It is the projection matrix of the linear space spanned by the cols of X  H = HT, HX = X, H2 = H, (I-H)2 = I-H, XTH = XT  Residual vector: **e** = **y** - = (I-H)**y**. Vector of fitted values: = H**y**  Hence **e**T**1** = 0 (by assumption 2), **e**T**x**j = 0 (by assumption 1), mean of = | | | | |
| Decomposition of Sum of Squares | | | | | Matrices HT, HR and HE are all symmetric and idempotent (i.e. HTH = H)  Hence SST = **y**THT**y** = **y**T(I – )**y**. SSR = **y**THR**y** = **y**T[X(XTX)-1XT – ]**y**. SSE = **y**THE**y** = **y**T[I – X(XTX)-1XT]**y**. SST = SSR + SSE | |
| Dist of Sum of Squares | | | Under assumptions of multiple regression models, (Note var(AY) = Avar(Y)AT, cov also similar)  Under hypothesis = ...= = 0, SSR and SSE are indep. ~ . ~ , | | | |
| ANOVA Table | | |  |  |  |  |  | | --- | --- | --- | --- | --- | | Source of variation | SS | df | MS | F-statistic | | Regression | SSR | p | MSR=SSR/p | MSR/MSE = Fp, 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 R2 = = = corr(**y**, )2  Adjusted R2: = R2 – (1 – R2) | | |
| 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 **z** of covariates (X1,...,Xp)T is MCORR(Y, **z**) = where **a** is a vector of constants, (i.e. **a**T**z** is LC of **z**)  = , where = (Cov(Y, X1),...,Cov(Y,Xp)), = variance matrix of **z** and = , = var(Y) | | | | |
| Let **y** be vector of n observations of y, Z be the matrix of observed **z** w ith row given by **z**i.  = =  In multiple LRM w vector of p covariates **z** = (X1,...,Xp)T, R2 = | | | | |
| Partial correlation coefficient | | Let X(-j) denote the sub-matrix of the design matrix X obtained by deleting col **x**j of X  Let be the residual of **y** regressed on X(-j), is residual of **x**j regressed on X(-j)  Correlation btw and is called partial correlation btw **y** and **x**j adjusting for effects of X(-j), given by  CORR(, ) = = where H-j is the hat matrix of X(-j)  Since residual of **y** regressed on X(-j) is part of **y** unexplained by X(-j), the squared partial correlation btw **y** and **x**j adjusting for effects of X(-j) is the proportion of unexplained variation of **y** which is explained by **x**j | | | | |
| Explicit expression of | | Explicit expression of can be obtained by considering the minimization of : 1st) minimize wrt with fixed, then minimize wrt , where is the sub-vector of eliminating | | | | |
| Minimizing w fixed : = | | | | |
| Minimizing wrt : = | | | | var() = |
| (Standardization) = , (which tells us which covariate contribute more to variation in y)  Note = CORR(, ) = = some constant \* | | | | |
| 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 is normal w mean = , var = (XTX)-1 | | | | |
| (n-p-1) ~ | | | | and are indep |
| Let = (XTX)-1. Let denote the jth diagonal elems of (estimated var of )  Note = cjj, where cjj is the jth diagonal elem of (XTX)-1  Then ~ tn-p-1; & For any constant vector **c**, ~ tn-p-1. | | | | |
| MSR and MSE are indep, since SSR and SSE are indep | | | | |
| Signifi-cance F-test | | Hypothesis: H0: = ... = = 0 vs H1: ≠ 0 for at least one of j = 1,...,p.  Test statistic: F = MSR/MSE. Under H0, F ~ Fp, n-p-1  For a significance level , reject H0 if F ≥ fp,n-p-1() or the p-value P(Fp,n-p-1 ≥ F) < ; otherwise, do not reject H0 | | | | |
| Wald test statistic | | Let be a vector of parameters, be its estimator, and the estimated var matrix of  Wald statistic for testing H0: = 0 is given by  Thus Wald test statistic for significance test H0: = 0, where = (,...,)T is given by W = , where is estimator of and is estimated variance matrix of  In context of multiple LRM, F = W/p, where p is dimension of | | | | |
| Individual t-test | | Answers qn: given other vars in model, does a particular predictor have a significant effect?  Hypotheses: H0 : = 0 vs H1: ≠ 0.  Test statistic: T = , where is the estimated SD of . Under H0, T ~ tn-p-1  For a significance level , reject H0 if |T| ≥ tn-p-1(/2) or p-value 2P(tn-p-1 ≥ |T|) < ; otherwise do not reject H0  - p-value better than test statistic as not only can reject H0, if p-value very small -> evidence supporting H1 is strong  - 1-sided test also same way as in simple LRM | | | | |
| Testing general linear hypothesis | | General linear hypothesis : H0 : = 0  For testing linear hypothesis, test statistic is T = . Under H0, t ~ tn-p-1  If only a few components of **c** are non-zero, T can be simplified.  E.g. **c** = (c1,c2,0,...,0)T, the var becomes var() + var() + 2c1c2cov(, ) and becomes c1 + c2  For a significance level , reject H0 if |T| ≥ tn-p-1(/2) or p-value 2P(tn-p-1 ≥ |T|) < ; otherwise do not reject H0 | | | | |
| CI and confidence bound | | A 100(1 - )% CI for is [ – tn-p-1(/2), + tn-p-1(/2)]  A 100(1 - )% CI for is  The 100(1 - )% confidence bounds for are | | | | |
| Prediction | | Given a new observation **x**0 = (1, x01,..., x0p)T, predicted value for both Ey0 and y0 is =  Estimated variance of fitted value is = (XTX)-1  Estimated prediction error variance is = +  Note y0 = + . But Ey0 = . Thats why prediction have extra error term  CI for Ey0 is ±  Prediction interval for y0 is ± | | | | |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| One-Way ANOVA | One-Way ANOVA model deals w only one factor, A having a levels.  , i = 1,...,a, j=1,...,ni, where yij is the value of Y for the jth member of the ith grp,  and are unknown params, are iid random errors, and ni is num of observations in ith grp  Let be mean response to ith grp. = . For identifiability, we restrict = 0.  Then represents the overall mean and represents effect of ith 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 = , SSA = , SSE =  = , = , n = | | | | | | | | |
| 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 = ... = = 0, F~Fa-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  , i = 1,...,a, j = 1,...,b, k = 1,...,nij, (gamma: interaction btw the 2 factors) where  = 0, = 0, = = 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 = | | | SSA = | | SSB = | | SSE = | |
| SSAB = | | | | = | = | | = | |
| = | | = | | = , | | n = | | = |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | 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: F1 = MSA/MSE, F2 = MSB/MSE. Under hypothesis of zero main effects, F1~Fa-1, n-ab, F2~Fb-1, n-ab  Factor w zero main effect factor has no effect, unless the interaction effect DNE | | | | | | | | |
| p-value for F1 = pf(F1, a-1, n-ab, lower.tail=FALSE). p-value for F2 = pf(F2, 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 denote expected response at level k.  Contrast is defined as , w = 0, i.e. **c**T**1** = 0, where **c** = (c1,...,ck) is called a contrast vector | | | | | | | |
| If factor has no effect, then all contrasts are 0, i.e. , 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: , .  If both not same -> there is interaction btw A and B.  Interaction effect is measured by | | | | | | | |
| In general, effect of factor A at a fixed level, i, of B is measured by any contrasts . If there is at least one pair (i,j) and at least one contrast **c**, s.t. ≠ , then A and B have interaction effect  Interaction effect is measured by interaction contrasts: (i.e. contrast of the contrast)  = , where , | | | | | | | |
| Interaction contrast vector is of the form (c1d1,...,c1db,...,cad1,...,cadb)  Components of the vector have restrictions: , j=1,...,b; , k=1,...,b;  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 | | Chart, line chart  Description automatically generatedChart, line chart  Description automatically generated | | | | | | | |
| Limitation of ANOVA | | Not convenient for detailed analysis of effects  In two-way ANOVA model, if nij (group sizes) are not the same, the SSAB does not measure the interaction effect, i.e.  in SSAB, the term 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: | | | | | | | |
| LRM for one-way ANOVA  Model: , expressed in dummy variables uk:  In matrix form: **y** = X +, where **y** = , = , X = . Alternatively, model can be expressed as  Regression params and mean response at ea level have relation: , k=2,...,a  Hence param (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. nij = 1; (ii) nij > 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: ,  Main effect model:  Expectation of the main effect model at diff levels:  = : diff of effect of A at level i and 1 when B is fixed at level 1, which is the same as , 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  = , diff of effects of B at level j and 1 when A is fixed at level 1, which is the same as , 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:  Expectation of the main effect model at diff levels:  = : diff of effect of A at level i and 1 when B is fixed at level 1,  = , diff of effects of B at level j and 1 when A is fixed at level 1,  = : diff of effect of A at level i and 1 when B is fixed at level j,  = , diff of effects of B at level j and 1 when A is fixed at level i,  = () – () = , 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. | | | | | | | |
| Inference on interaction effect | | For testing whether there is a significant overall interaction effect, under the regression model,  H0: = 0, i=2,...,a; j=2,...,b; vs H1: at least one of ≠ 0, where 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 = , where is vector of estimated 's and is the estimated covariance matrix of . F-statistic is F =  Under H0 F~F(a-1)(b-1), n-ab. Reject H0 at level , if F > F(a-1)(b-1), n-ab () or p-value pf(F, (a-1)(b-1), n-ab, lower.tail=FALSE) < | | | | | | | |
| For testing of particular interaction effect: perform individual test on 's and on linear combination of '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. H0 : () – () = = 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. H0 : () – () = – = 0  3) Test whether diff btw level j and k of B is same at level i and l of Factor A, i.e. H0 : () – () = – – + = 0  General rule of thumb: if subscript of contains 1: ignore; else convert to  Since any particular interaction contrast is a LC of 's, i.e. **c**Twhere is the vector of . Let be the vector of estimated 's and is the estimated covariance matrix of  Test statistic for **c**T= 0 is = . Under H0: **c**T= 0, ~tn-ab  E.g. for a particular interaction contrast, above formula can be simplified. H0: – – + = 0. Only vector = and its covariance matrix are needed. And corresponding **c** reduces to **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 from main-effect model not correct estimate of error variance. SSEM = SSEI + SSAB (error caused by iteraction + interaction effect). Instead, from iteraction model has correct variance | | | | | | | |
|  | | Let Xm denote design matrix of main-effect model. The estimated var matrix from main-effect model is V.m =  It shld be adjusted to V = = 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 . For any j , let Tj be test statistic & decision rule to reject H0 be Tj ≥ cj for a critical value cj  Family-wise type I error rate is P() | | | |
| 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 be the vector of the grp means. Let **c** denot a contrast vector of and denote set of all possible contrasts.  Then H0: **c**T= 0 for all **c** . H0 can be expressed in terms of the basis contrasts:  Let be the vector of basis contrasts. H0 is then equivalent to **a**T = 0 for any **a**  In main effects: = (, ..., ) or (,..., ). In interaction effects: = ()i=2,...,a; j=2,...,b | | | |
| Scheffe's soln  For an individual contrast **a**T, test statistic T**a** = , where is estimate of , is estimated var matrix of  Need to find critical value s.t. P() ≤  To control family-wise type I error rate at level , Scheffe's soln is to take for all j, cj = = , where m1 is num of components of and m2 is df of , is the upper quantile of the dist. ( 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 is and = . Note =  Let **b** = , **c** = (c1,...,ca)T, D = diag(). Then = = **b**TD-1**b** = SSA =  And = (a-1)MSA/MSE = SSA/MSE (since SSA = (a-1)MSA, = MSE) (where MSA/MSE is )  p-value = P( F-ratio) | | | |
| Then the simultaneous p-value of 2-sided test is P() = P() = P()  For 1-sided test: p-value is (1/2)P() and critical value is = , | | | |
| Individual & Simultaneous CI | | | | Simultaneous CI: CI which covers all params  Simultaneous CI w confidence coefficient 1- for all j require that P() ≥ 1-  Simulateous CI for the contrasts is ± |
| 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 in summary data or look at estimated regression coefficients) | | | |
| Multiple comparison through LRM | LRM for one-way ANOVA is , i = 1,..,n where = , k ≥ 2  For any contrast, we have = = =  Test statistic: TC = , where = (c2,...,ca)T, is estimated covariance matrix of (,...,)  TC is same as TC when using level means | | | |
| Pairwise contrasts | Only interested in certain pairwise contrasts, , 1 ≤ k < j ≤ a. Overall significance F-test not necessary  Only need to find s.t P = | | | |
| Studentized range dist | Let , i=1,...,a be sample means of a samples w equal sizes (ni = n)  Studentized range statistic is =  Let denote upper -quantile of the Studentized range dist, aka Tukey's criterion for pairwise comparison at level | | | |
| Pairwise comparison procedure | For studentized range dist, Q-statistic for is Qij = , where  Contrast is significant at level if Qij > | | | |
| Diff btw Q-statistic & t-statistic | So |Tij| = Qij/  t-statistics can be used for pairwise comparison using Tukey's criterion, but |Tij| must be compared w  i.e. contrast is significant at level if Qij > OR |Tij| > | | | |
| Equivalent form in terms of regression coefficient | | | W conventional defn of dummy vars, the regression coefficients = , i = 2,...,a OR = , i,j > 1  t-statistics: Tij = | |
| Tukey's simultaneous CI | | | For = : ± sd()  For = , i,j > 1: () ± | |
| Bonferroni's Mtd | | If there are only k contrasts we are interest in, the overall type I error rate for the k contrasts can be controlled by Bonferroni's Mtd: , where is type I error rate for contrast j  Each can be specified, but in general just use = /k. Critical value = Tn-a(/k\*2). Check |Tj| > for 2-sided test | | |
| Rationale: P() ≤ , where P() ≤ ≤ . Thus Bonferroni mtd is a conservative mtd | | |
| p-values: p-value for jth test is pj = kP(Tj ≥ ) where is observed value of Tj and prob computed under dist of Tj  If critical value for all individual test is set at , then = P(Tj ≥ ) for j = 1,...,k and overall type I error rate is = k  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,  where uj, j = 2,...,l are dummy variables representing a factor predictor and x is a quantitative predictor, is diff of Y btw type j and type 1 when X is same for both types, is effect of X on Y (i.e. when X change by a unit, Y has an expected change of )  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 = , i.e.  Category 1: Y = . Category i ≥ 2: . So just making inference on , and | |
| LRM w non-linear predictor terms | | | | In LRM: Y = + X1 +...+ Xp + , X need not be diff predictor variables, could be non-linear fns of predictor variables  E.g. Y = +Z+ Z2 +...+ Zp + | Y = + (1/X) + (inverse model) | log(Y) = + log(x) + log(v)+ (log model) |
| Polynomial regression models | | | | yi = + x1i + x2i + + + , i = 1,...,n (model can have more than 1 predictor variables)  yi = + + + + + (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: = (X - c)+ =  Model from yi = + Xi + becomes yi = + Xi + + (piece-wise linear in X)  Slope of model changes from to 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 = (X1, X2) and includes all relevant predictors.  **y** = X + = (full model). **y** = (reduced model)  Under full model, , E = , Var() = . Var() =  Under reduced model, = , Var() = , E =  Under reduced model, LSE of is larger and have a bias of  Under reduced model, var of LSE of is smaller since > (then inverse it to get < )  On the other hand, over-fitting increase variance, but decrease bias | |
| Effect of under/over fitting on prediction | | Let be the parameter of model M w design matrix XM. The estimator of E**y** = under model M is  and E and Var() =  , where M is num of cols of XM (usually num of predictors in model M + 1)  If we use model M to predict n unobserved y's w the same XM, the sum of prediction squared error (SPSE) is SPSE =  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 is measured by MSE which also consists of a variance and bias component. MSE = = var() + bias2()  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 | |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | |  | R2 and | Mallow's Cp | AIC | BIC | EBIC | CV | | Suitable for | models w same num of predictors | Good estimate of | Minimize Kullback-Leibler dist |  | models in small-n-large-p problems | based on estimated prediction error | | Better | Higher | Lower | Lower | Lower | Lower | Lower | | | |
| R2 and | | R2 = , = R2 – (1 – R2) both estimate proportion of pop variation of response Y explained by predictors in regression model  However, R2 always incr when num of predictors incr. Does not strike a balance btw variance and bias  Though 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 Cp (Complexity parameter) | | Ideally, use to estimate SPSE. But don't have yn+i available. So use to estimate SPSE and find that  E = SPSE - 2|M|  Adjusting for bias, (unbiased estimate of SPSE)  Mallow's Cp = . Minimizing Cp is equivalent to minimizing  If model M is correct model, then ECp ≈ |M| | |
| Akaike's information criterion (AIC) | | AIC minimizes the Kullback-Leibler dist btw model M (gM(y)) and the true model (f(y) approximated by the empirical dist) given by  I(f,gM) =  AIC approximates the 2nd component above (since can ignore constant = 1st term)  AIC = , where are MLE under model M, L is likelihood fn and jM is num of predictors in M  For multiple LRM w normality assumption, AIC = , where and C = n(ln(2π) + 1) | |
| Bayesian information criterion (BIC) | | Let be the set of all possible models, Pr a prior probability measure on . Let Pr(M) be the prior on a model M and the prior on parameters of model M  The marginal density of data **y** given model M is m(**y**|M) =  The posterior probability of model M is p(M|**y**) =  If Pr(M) is taken as constant, BIC = | |
| Extended Bayesian information criterion (EBIC) | | Let model space be partitioned according to num of predictors contained in the models as , where is set of all models containing j predictors. If Pr is the constant prior, then Pr() , 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 = , | |
| Cross Validation (CV) | | Model M fitted using data (**y**, X). To assess goodness of model, ideally use another data set, () and validate by prediction error . 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 = , where is the estimate of by leaving out the ith data point (yi, **x**i)  k-fold CV score: CVk = , where (**y**j, Xj) is jth part of data (testing data) & = estimate obtained by training data | |
| Model selection strategy | | Naive method: all subset selection. Not practical as for p predictors, there are 2p 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  ii. Fit full model, remove predictor w largest p-value which is > , then re-fit model w remaining predictors, repeat this until no predictor  has p-value >  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 M0 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 R2 and | |
| 3) Backward Selection (inverted sequential method)  Start w full model MF 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 **y** = X+ , the penalized likelihood approach select variables by minimizing , where is the penalty function, is the penalty parameter whose value is to be chosen  Procedure: specify sequence of 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:  2) Adaptive LASSO penalty: , where wj is taken as . If p n, being the OLS (ordinary least square) estimator in multiple LRM. If p is close to or > n, is the OLS estimate is the marginal LRM  SCAD penalty: for near 0, and equals a constant C for large , the two parts are connected by a smooth function  MCP penalty: for large , it is a constant. Smoothly decreases to 0 w = as its asymptote when approaches 0 | |
| Rationale of penalized likelihood approach | | E.g. LASSO (least absolute shrinkage and selection operator) estimates by minimizing  If = 0, LASSO estimator is same as LSE. If = ∞, all components of are estimated as 0  For a certain nonzero , 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 , i = 1,...,n. Hat matrix is H = X(XTX)-1XT  Hat values (ith diagonal elem of H), = hat value of ith observation, where is the ith row of X | | | | | | | | |
| Partitioning X as X = (**1** Z), (XTX)-1 =  A22 = = sample covariance matrix, A11 = , A12 = , A21 =  Let ith row vector **x**i of X be and . Then A11 =  Then hii = = = + Mahalanobis dist of ith observation in x-space to centroid of sample | | | | | | | | |
| Pearson's residuals are ei = yi – , i = 1,...,n. Let **e** = (e1,...,en)T. Then **e** = **y** – = (I – H)**y**. E**e** = **0**, var(**e**) = (I – H). var(ei) = (I – hii)  Under normality assumption, **e** ~ N(**0**, (I – H)) | | | | | | | | |
| Studentized residuals,  Studentized deleted residuals, , where is the predicted value by fitted model w ith observation deleted, is the counter part of when ith observation is deleted  Cook's distance, di = , where is estimate of when ith observation is removed from data  Variance Inflation Factor (VIF). Let be the coefficient of determination of model Xk = . VIF of Xk is VIFk = | | | | | | | | |
| 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 | | | | | | | Chart  Description automatically generated Chart, line chart  Description automatically generated | |
| 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 yq and zq denote q-quantile of Y and Z, then P(Y ≤ yq) = P(Z ≤ zq) ≥ P(Y ≤ zq). Then P(Y ≤ yq) ≥ P(Y ≤ zq)  Hence yq ≥ zq for all q. In Q-Q plot where yq is plotted against zq, the point (zq, yq) is above the point (zq, zq) | | | | | | | | Chart, line chart  Description automatically generated |
| 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 = Tr(H) = p. Point is high leverage if hii > | | | | | | | | |  |  |  |  | | --- | --- | --- | --- | |  | a | b | c | | Leverage | low | high | high | | Consistency | No | Yes | No | | Influence | low | low | high | | |
| 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 || 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 h­i and Cook's distance di  Formal test: To assess (yi, **x**i), introduce the dummy variable u = .  Significance of coefficient of u in the linear predictor indicates ith point is an outlier. | | | | | | | |

|  |  |  |
| --- | --- | --- |
| 1. LRM w unequal variances | When 's don't have common variance, an unequal variance model is considered: ,  wherehas the variance matrix , where wi's are unequal weights (if common variance: ) | |
| If wi's are known, the unequal variance model can be transformed into an equal variance model. Let  W = . Multiply by W1/2, then **.** Let **,** ,  Then var() . And model is a constant variance model | |
| Minimizing , we obtain the estimate of as = weighted LSE (WLSE)  can be expressed explicitly as . The weight wi reflect the relative importance of in the estimation. The larger the variance of ith term, the smaller the corresponding weight, since weight is inversly proportional to the variance | |
| For WLSE , E = , var() = . And  is estimated as . Inference on 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  2. Regress log of absolute residual on log of fitted values, e.g. ln |ri| = .  If smallest < 0, replace by , for some positive constant c  3. Weights are estimated as OR . Constant 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 si = , where si is sample sd, and is sample mean  2. The weight for the ith predictor value is | |
| 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, XTX would be singular, i.e. non-invertible  Practically, although perfect linear dependence will not occur, high multicollinearity can cause XTX to be nearly singular (i.e. having large condition number ), 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 | being the submatrix of X w/o its jthcolumn and is the coefficient of multiple determination of **x**j is regressed on  If **x**j is uncorrelated w -> = 0 and the variance of = /SSTj  If **x**j is correlated w -> the variance is inflated by a factor VIFj = (Variance inflation factor) | |
| 2) Ridge regression | Since consequence of multicollinearity is XTX nearly singular, to remedy: add diagonal matrix to XTX, i.e. XTX + I (which is invertible)  Ridge regression estimator , where > 0 is a parameter to be chosen  is also the minimizer of the penalized sum of squares: | |
| E = -> estimator is biased. Bias is  Let Q be the orthogonal matrix s.t. XTX = QQT, where = Diag(,...,). Thus QT  Let denote the jth component of QT. We have . Thus bias increase as increase | |
| Thus tr(var()) = . The variance decrease as increase. All other cov = 0 | |
| Sum of mean squares errors of is given by MSE = .  Need to get balance btw bias and variance by minimizing MSE  MSE cannot be readily used as a criterion as it involves unknowns and . Can select by Cross validation (CV)  Let be the ridge regression estimate of w parameter by deleting the ith observations. The CV score is given by CV() = , where is the ith row vector of the design matrix X  Best is the minimizer of CV() | |
| 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 transfor-mation) | 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 depend on mean , i.e. = V(), 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 , as h(Y) ≈ h() + h'()(Y-), where h() is a constant  Treating h() as mean of h(Y), var(h(Y)) = E[h(Y) - h()]2 ≈ [h'()]2E(Y - )2 = [h'()]2V()  WLOG, setting [h'()]2V() = 1, h'() = -> h() = , aka Variance Stabilization transformation | |
| Box-Cox transfor-mation | 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) = . can be determined by data  When = 0, the Box-Cox transformation is given by h(Y) = ln(Y), since = ln(Y) | |
| Let , where and are the SD and mean of ith treatment effect. The in Box-Cox transformation is determined by variance stabilization transformation   |  |  |  |  | | --- | --- | --- | --- | |  | (just use closest half int, e.g. = 0.04 -> use 0.05) | = 1 - | Transformation | |  | 3 | -2 | reciprocal squared | |  | 2 | -1 | reciprocal | |  | 3/2 | -1/2 | reciprocal sqrt | |  | 1 | 0 | log | |  | 1/2 | 1/2 | sqrt | | constant | 0 | 1 | no transformation | |  | -1/2 | 3/2 | 3/2 power | |  | -1 | 2 | square | | |
| Determination of  Method 1: If observations are grouped, for each group, compute and . Fit regression model ln si = + ln + ei  If observations not group, fit ln |ri| = + ln + ei  For both, = estimate of  Method 2: Only for grouped observations. Select a few values, say , k = 1,...,K. For each k, compute Rk =  Select with smallest Rk | |
| Direct determination of  For grouped observations: select a few values, say , k = 1,...,K  For each k, make the transformation yij . With the transformed data, compute , i = 1,...,g  Select s.t is closest to 1  For non-grouped observations: select a few values, say , k = 1,...,K  For each k, make the transformation yij . Analyze the regression models w as response variable.  Select s.t MSE() is smallest | |