Response Variable: random predicting variable: fixed

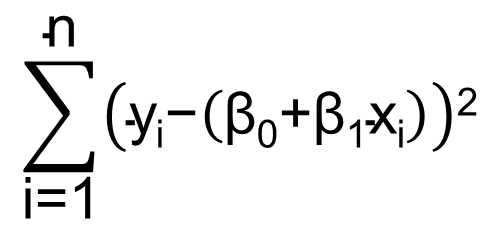
Regression: Prediction Modelling Testing

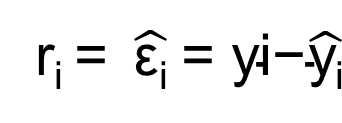
is the deviance of the data from the linear model （normal distribution）

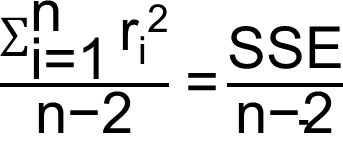
Assumptions:

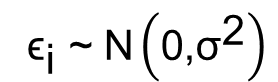
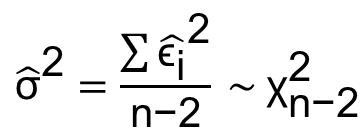
* *Linearity/Mean Zero Assumption :* E(i) = 0
* *Constant* *Variance Assumption*: Var(i) = 2
* *Independence* *Assumption* {1,…, n} *are independent random variables*
* (*Later we assume* i ~ *Normal)*

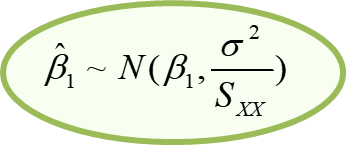
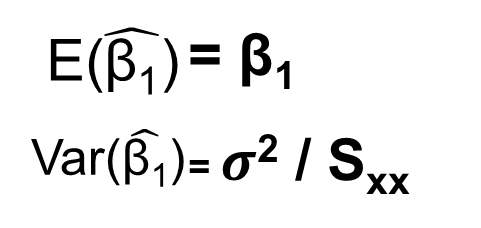
model parameters are: 0 1 2 (unknown)

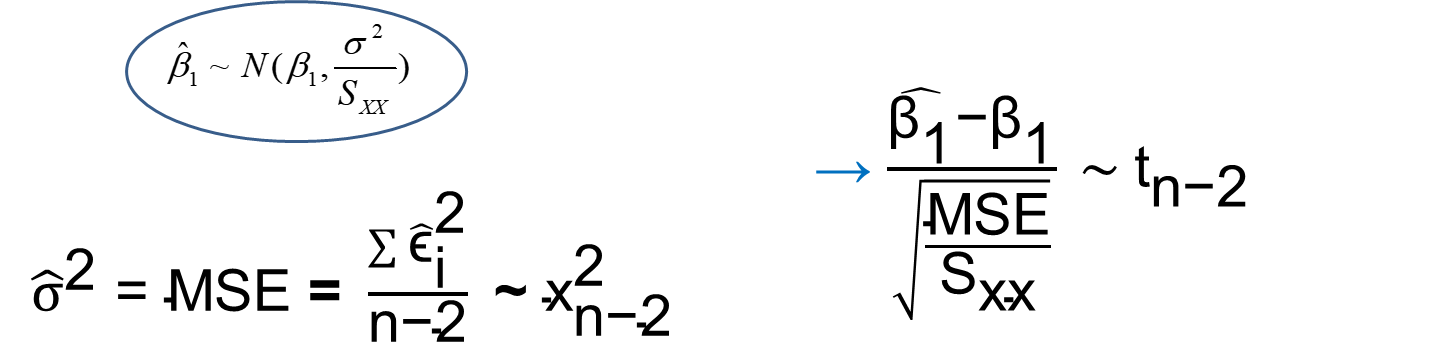
Minimize 

Residuals: 

*Mean squared error*: Estimator for 2 （Chi squared distri）:

>> sampling distribution of



Sampling Distribution of

The estimators for the regression coefficients are Unbiased regardless of the distribution of the data.

The assumption of normality s needed for the sampling distribution of the estimators of the regression coefficients and hence for inference.

is normally distributed

prediction contains two sources of uncertainty: new (n+1)observ, parameter estimates (of b0 and b1 )

To model the nonlinear relationship, we can transform X by some nonlinear function

Normality assumption does not hold: Transform the response variable from y to y\* via y\* = yl

R2 = 1 – SSE / SST

ANOVA: *Comparing means from multiple populations assuming the variances are the same*

*and equal to . DF: N-k*

The sampling distribution of the pooled variance is a chi-square distribution with N-k degrees of freedom.

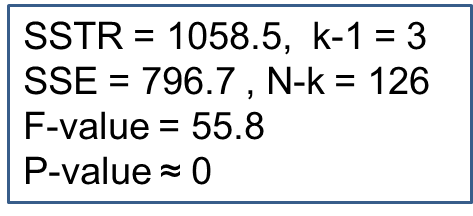
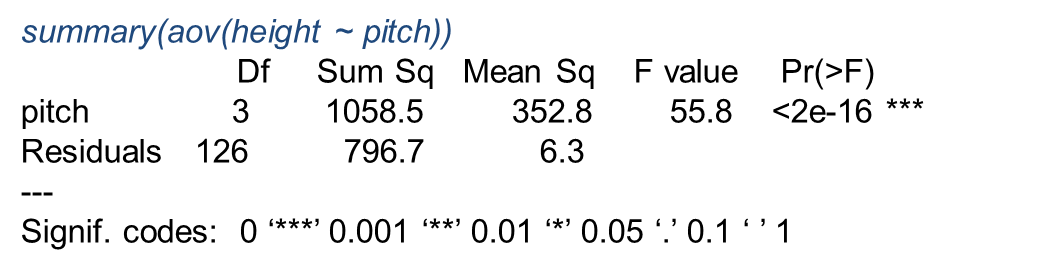
SST = SSE + SSTR,

MSE = SSE/N-k = *within-group variability*

MSSTR = SSTR /k-1 = *between-group variability*

ANOVA: comparing between to within variability

F = between-group variability/within-group variability = F(k-1,N-k)

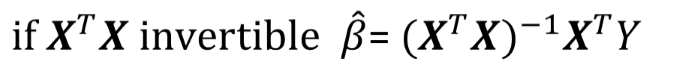


Q: “studentized range” distribution. *q* > *t* (at any fixed and df)

Assumptions:

* *Constant* *Variance* *Assumption*: Var() = 2
* *Independence* *Assumption:* {1j,…, nj} are independent random variables
* *Normality* *Assumption*: ~ Normal(0, 2)

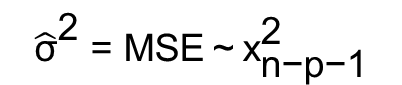
ANOVA: A linear regression model where the predicting factor is a categorical variable.



The estimator of 2  is MSE:

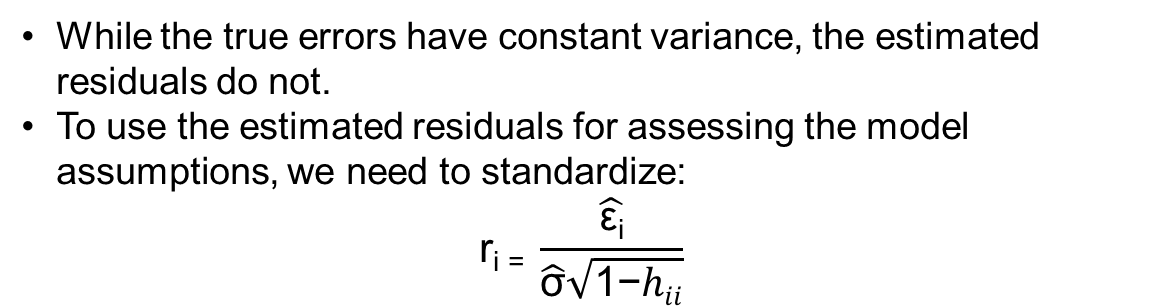
* Assuming e1,…, en  are normally distributed, them

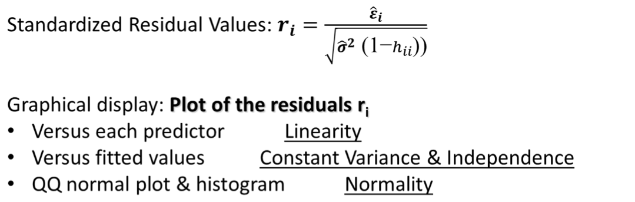
MSE ~ x2 with n-p-1 degrees of freedom (Why n-p-1?)

include intercept. sampling distribution of

 is a linear combination of {Y1,…,Yn}. ei ~ N (0, 2), sampling distribution of is normal. CI for tn-p-1

Sampling distribution of Y is tn-p-1





F0 = MSR / MSE ~ F( k, n-k-1 )

*MSR = SSR/k MSE = SSE/n-k-1*

Multiple regression: full model---conditional model

interpret the statistical significance in a multiple regression model conditionally

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | DF | Sum of Sq | Mean SS | F-statistic |
| Regression | P | SSReg | SSReg/p | MSSReg/MSE |
| Residual | n-p-1 | SSE | SSE/n-p-1 |  |
| Total | n-1 | SST |  |  |

SSE(x1) = SSR(X2|x1) + SSE(X1,X2)