**fariSample mean**

**Central Limit Thm**:

**Sample Variance:** ,

Use to make E[sample var] = population var.

**Chi Square** Use to find from normal.

are iid standard normal

, is chi-square with deg freedom

**Sampling from a Normal Population**:

,

**T-Distribution** If , ,

**Method of Moments**

If unknown parameters, let

**Maximum Likelihood Estimator**

Likelihood

Maximize the log likelihood

**Estimate Difference in Means**

Get pooled sample variance:

Like a weighted average, weights as degrees of freedom

**Evaluating a point estimator**

Biased: shooting off-target

Mean Squared Error:

=

Hard to assess credibility of a point estimate by itself

Precision: measure spread

**Confidence Interval**

Step 1: find a statistic involving the parameter

Mean with known variance:

Mean with unknown Var:

Var with unknown mean:

Step 2: find initial interval with prob.

Two-sided:

Step 3: Find the equivalent interval for the parameter

Step 4: plug in observed numbers

One-sided upper:

known, use

unknown but equal use

*Lower confidence interval*

known, use

unknown but equal use

*Approximate Confidence Interval*

Have CI with length no greater than . What’s the min sample size required?

Length:

Sub with after testing times

**Hypothesis Testing**

: presume this, in control, conventional, not effective, no discrimination­­­

out of control, new finding, discrimination, more effective

Level of significance *Critical Region Approach*

Step 1: set up null and alternative hypotheses

Step 2: find a statistic with known distribution

Called the test statistic TS,

Step 3: Choose an and find “blue region”

Smaller means smaller region

Step 4: check if observed TS falls into region

*Critical Region* is set of all samples that will lead to rejection of

*P-value* Prob of observing evidence as much/more in favor of

Step 1: Set up null and alt hypotheses

Step 2: Find TS with known dist under

Step 3: Calculate p-value

Step 4: check if is small enough

If is smaller than your , reject

tells us how cautious/easily convinced we should be

*Decision Errors* Type I: Reject when it’s actually true

“False alarm” error

Type II: Accept when it’s not true. Miss.

Operating characteristic (OC) curve:

is Power Function

*One-Sided Test*

TS: ,

*Alternately* , composite hypothesis

TS same as above,

(the bell curve only has one region that matters)

*Unknown Variance* , TS

*Equality of Means* and are known, are normal

2 sets of indep samples

TS: if is more likely to be small.

since sum of indep normal is still normal.

Standardize: =TS.

= and reject if is small

*Unknown Variance* Assumed equal

similar for

Pooled sample variance as weighted average

If is true, then

*If Assume not equal* replace with and

*Test Equality of Proportions* L18

**Linear Regression**

Sum of squared residuals:

Choose and that minimize this

Take partial derivatives w/ respect to a, b

Set to 0, get normal equations ,

Solution: ,

Identities:

So

This is TSS, total sum of squares, amount of variance in response variables. Part of this is caused by difference in input variables

Measures remaining amount of variance in the response variables, after we take into account difference of input variables, known as RSS, residual sum of squares

is amount of variation explained by input var, known as explained sum of squares ESS

Coefficient of determination

This shows how well the model fits the data.

Ex. , so 75% of variation of y is from changes in x. The rest is from variance even when x is constant

So

Sample correlation coefficient:

So , traditionally

does not measure the slope of regression line

Slope is , is about if

so usually you get regression to the mean

**Inferential Linear Regression**

where is error, assume

For each observation

Assume random errors are indep

Estimate from the tuples of samples

Estimate and from least squares formulas

estimates

,

is normally distributed, unbiased estimator for

, lose 2 deg. Free since A and B are linear combinations of , , . indep of A and B

*Confidence Intervals*

:

:

:

*Distributional Results*

:

:

Mean Response :

Future Response :

*Mean Response given :*

Estimates mean response given input

Unbiased, is a linear combination of indep normal r.v.

*Prediction Interval*

Find an interval that will contain the response with certain confidence, NOT the same as confidence interval

Best “point prediction” , unbiased

**Analysis of Residuals**

Assume independence, constant variance (homoscedasticity), normality.

Check these assumptions by looking at residuals

Standardized residuals (the denominator estimates

Linearity violated if residuals show discernible pattern

Nonconstant variance shows increasing outliers

*Check Normality*

Sample quantile of standardized residuals vs standard normal quantile

Quantiles: points taken at regular intervals from CDF of r.v.

Q-Q plot: compare 2 distributions by plotting quantiles against each other, to check normality

*Transforming to Linearity*

Where is and is and is Y

Interpret the regression coefficients:

Regression equation

average number of defects at is , and as X increases by 1, number of defects increases by on average

by Taylor expansion

by Taylor expansion

*Linear Transformations*

Exponential:

X increases by 1, Y changes by factor of 100d%

so where

Logarithm:

X increases b factor of 1%, Y changes by d/100

so

Power:

X increases by factor of 1%, Y changes by factor of d%

,

**Multiple Linear Regression**

*Least Squares method:*

where and is amount of change in response if increased by 1 keeping others unchanged.

Residuals

Matrix form

residual sum of squares

Coefficient of multiple determination

proportion of variation explained by model

is multiple correlation coefficients

As input vars increase, goes down, goes up

…But this could mean overfitting

Adjusted/corrected

*Inference* Has multivariate normal dist , is unbiased estimator

is covariance matrix, diagonals give variance

For each

and is elem in diagonal of

and independent of B

This is an unbiased estimator for

for inference on

and denominator is estimate for sd of , standard error. Test whether has impact on response keeping other input vars unchanged

*Reading Excel*

Multiple R:

R Squared:

Adjusted R Squared:

Standard Error:

Observations:

**Polynomial Regression** ,

**ANOVA:** Analysis Of Variance

Study the impact of a certain factor

: all means are equal

total number of observation

sample mean of all observation

th observation in th group

Sample mean of the th group

Between-group sum of squares

Amount of variation that can be explained by groups

Within-group sum of squares

amount of var that can’t be explained by groups, , and are indep.

TS: if is true

-value=

*ANOVA table for multiple regression* Y doesn’t depend on input

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| source of var | sum of squares | d.f. | mean square | F ratio | p-value |
| between groups |  |  |  |  |  |
| within groups |  |  |  |  |  |
| total |  |  |  |  |  |

**Two-Factor ANOVA**

One-way: single factor

Ex. gas mileage: difference between cars and with drivers?

Each row/column represents one level

Column averages

Row averages , overall average

*Assumptions* are independent normal r.v.

Variances of each are the same

where is overall mean

is correction for the th row

is correction for the th column

is mean for row, use for mean of column

*Hypotheses* row factor has no impact, : colun has no impact

*Estimators* is an unbiased estimator for , overall sample mean overall mean

is unbiased estimator for

is unbiased estimator for

Sample mean difference mean difference

Row sum of squares

Variation between rows

Column sum of squares

Variation between columns

Error sum of squares Variation due to randomness, subtract mean, and from

an unbiased estimator for

Error:

Row: if is true, , indep. Of

Col: is true, , indep. Of

*ANOVA Table* Let

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| source of var | sum of squares | d.f. | mean square | F ratio | p-value |
| row |  |  |  |  |  |
| column |  |  |  |  |  |
| Error |  |  |  |  |  |
| total |  |  |  |  |  |

**2-way ANOVA with Interaction**

Without interaction, is same for all

same for all

*Interaction*

: grand mean

effect of the ith row

effect of the jth col

interaction of ith row and jth column

*Data* m rows, n columns, m\*n cells, 2 observations/cell : sample mean of cell (,j)

denotes an individual number

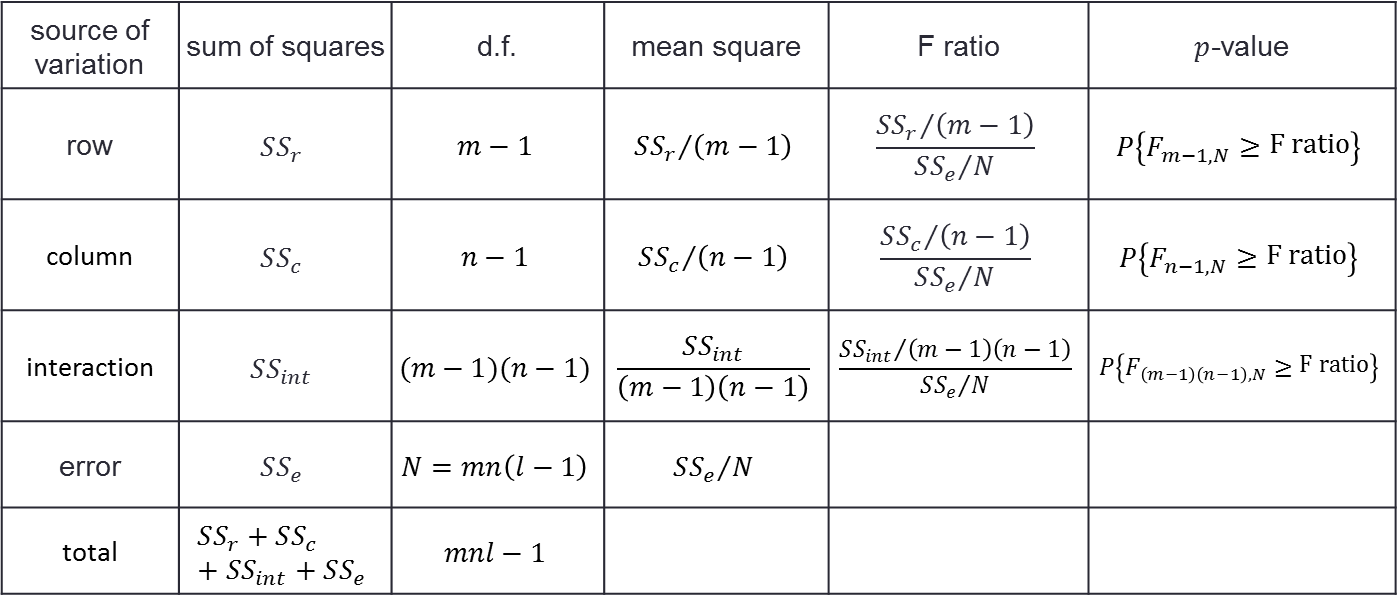
*Assumptions*

Independent random error,

: no {row,column,interaction}

*Estimation*

* + Row effect
  + estimator:
  + Column effect
  + estimator:
  + Interaction
  + estimator:
  + Error estimator
  + Error sum of squares
  + , is an unbiased estimator for
  + Interaction sum of squares , Under ,
  + Row sum of squares
  + Under ,
  + Column sum of squares
  + Under ,



**Sample Midterm 2**

(a) If simple linear regression of Y on X yields a line with slope equal to 2, then simple linear regression of X on Y will yield a line with slope equal to 1/2. False.

(b) A simple linear regression model (Y on x) has R^2 = 0.7, while an exponential

model (log(Y) on x) has R^2 = 0.8. This means the exponential model can explain more variation in Y. False.

(c) In multiple regression, adding an additional regressor (input variable) will always increase the multiple R-square, and decrease the adjusted R-square. False.

(d) In linear regression, a 95% prediction interval for a future response given certain

input level always contains the 95% confidence interval for the mean response given the same input level. True.

(e) In one-way ANOVA, the null hypothesis states that the population mean within each group is equal to zero. False.

Having done poorly on their math final exams in June, some students repeat the course in summer school, then take another exam in August. Assume these students are representative of all students who might attend summer school. Given their scores in June and August, explain how you can test whether the summer program is worthwhile.

Use paired t-test. is mean score in June, mean score in August: not effective

Let Xi and Yi be the scores of ith student, and take the difference Di=Yi-Xi. Sample mean , and sample sd

TS: is

3. In a study of the relationship between the starting salary (Y) and GMAT score (x)

of n = 100 MBA students, the following results are obtained: average starting salary

avg(Y) = $90,000, sample standard deviation s(Y) = $45,000, average GMAT avg(x) = 600,

sample standard deviation s(x) = 100, sample correlation coefficient r = 0.3.

(a) (10 pts) Find and interpret the slope of the simple linear regression of Y on x.

=90000-135\*60=9000

Y=9000+135x

(b) (10 pts) Find a 95% confidence interval for the mean salary of people who got 700 in GMAT.

Get 100 C.I. for

Use

(c) (10 pts) Test at 0.05 level on whether higher GMAT score means higher salary.

mean salary doesn’t increase GMAT

TS is ,

p-value =

Since close to standard normal, know p=1-

(d) (15 pts) Construct the ANOVA table for linear regression.

(e) (10 pts) Simple linear regression of log(Y) on x yields slope = 0.015%, and that

of Y on log(x) yields slope = 8000. Interpret the results.

If GMAT incrases by 1, starting salary increases by .015% factor on average.

For slope=8000: if GMAT score increases by factor of 1%, salary increases by 8000\*.01=80 on average.

An emergency room physician wanted to know whether there were any differences in the amount of time it takes for three different inhaled steroids to clear a mild asthmatic attack. Over a period of weeks she randomly administered these steroids to asthma sufferers, and noted the time it took for the patients’ lungs to become clear. Afterward, she discovered that 12 patients had been treated with each type of steroid, with the following sample means (in minutes) and sample variances resulting. Construct the ANOVA table to test on whether the mean time to clear a mild asthma attack is the same for all three steroids.

Steroid Sample mean Sample variance

A 32 145

B 40 138

C 30 150

=3, n=12, is mean time to clear attack for ith steroid

: all means are equal,

=

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| source of variance | sum of squares | d.f. | mean square | F ratio | P-val |
| between groups | 672 | 2 | 336 |  |  |
| within groups |  | 33 | 144.3333 |  |  |
| total | 672+4763=5435 | 35 |  |  |  |