Probability and Statistics Notes

Kari Innes

September 12, 2014

## Data Collection

Classify a study as observational or experimental, and determine whether the study’s results can be generalized to the population and whether they suggest correlation or causation.  
- If **random sampling** has been employed in data collection, the results should be generalizable to the target population.  
- If **random assignment** has been employed in study design, the results suggest causality.

Identify the explanatory variable in a pair of variables as the variable suspected of affecting the other, however note that labeling variables as explanatory and response does not guarantee that the relationship between the two is actually causal, even if there is an association identified between the two variables.

An extraneous variable that is related to the explanatory and response variables and that prevents us from deducing causal relationships based on observational studies is called a confounding variable.

Simple random sample - each case is equally likely to be selected

Stratified sample - divide the population into homogenous strata, then randomly sample from within each stratum.

Cluster sample - divide the population clusters, randomly sample a few clusters then randomly sample from within these clusters

Four principles of experimental design  
-control any possible confounders,  
-randomize into treatment and control groups,  
-replicate by using a sufficiently large sample or repeating the experiment,  
-block any variables that might influence the response.

Explanatory variables (factors) - conditions we can impose on experimental units

Blocking variables - characteristics that the experimental units come with, that we would like to control for

Blocking is like stratifying: -blocking during random assignment -stratifying during random sampling

Explanatory and response variables can show correlation but not necessarily causation.

Note that there are three commonly used measures of center and spread:  
- center: mean (the arithmetic average), median (the midpoint), mode (the most frequent observation)  
- spread: standard deviation (variability around the mean), range (max-min), interquartile range (middle 50% of the distribution)

A robust statistic (e.g. median, IQR) as a statistics that is not heavily affected by skewness and extreme outliers.

Left skewed distribution: mean is lower than median and mean is closer to the upper bound of the distribution

Right skewed distribution: mean is higher than median and mean is closer to the lower bound of the distribution

Unimodal, bimodal and multimodal distributions:  
-Box plots do not provide any information about the modality of a distribution. Boxplot shows the inner quartile (Q1 to Q3 or 25% to 75%), outliers and median. Box to left of page shows right skew, box in middle shows symmetric and box to right of page shows left skew.

Barplots for categorical variables, histograms for numerical variables

Use contingency tables and segmented bar plots or mosaic plots to assess the relationship between two categorical variables.

Law of large numbers: as more observations are collected, the proportion of occurrences with a particular outcome converges to the probability of that outcome

## Probability and Distributions

### Disjoint vs. Independent

-Disjoint (mutually exclusive) events as events that cannot both happen at the same time: If A and B are disjoint, P(A and B) = 0 and P(A or B) = P(A) + P(B).  
-If A and B are independent, then having information on A does not tell us anything about B (and vice versa).  
-If A and B are disjoint, then knowing that A occurs tells us that B cannot occur (and vice versa).  
-Disjoint (mutually exclusive) events are always dependent since if one event occurs we know the other one cannot.

For non-disjoint events, P(A or B) = P(A) + P(B) - P(A and B)

If P(A | B) = P(A), then A and B are independent

Probability distribution: -the events listed must be disjoint -each probability must be between 0 and 1 -the probabilities must total 1

Complementary events are two mutually exclusive events whose probabilities add up to 1.

The probabilities of two disjoint outcomes may not always add up to 1 because there may be more than two outcomes. The probabilities of two complementary outcomes always add up to 1.

Posterior probability is generally defined as P(hypothesis | data). This is different than the p-value which is the probability of observed or more extreme data given the null hypothesis being true, i.e. P(data | hypothesis)

### Test Statistics

SE: Standard Error - variability in point estimate from different samples of same size from the same population. Equals sd / sqrt(n)

sd: Standard Deviation - variability of sample data. Equals sum of (difference between each point from mean) ^ 2. This is squared to get ri of negatives and to give greater weight to larger differences.

sigma: Variablitiy of population

In R, qnorm gives the critical value for an input confidence level

qnorm(.975, mean = 0, sd = 1)

## [1] 1.96

pnorm gives the probabilty of being greater than an input critical value (for given mean and standard deviation)

pnorm(.8, mean = 0, sd = 1)

## [1] 0.7881

Applet to use for checking r results: <http://spark.rstudio.com/minebocek/dist_calc/>

### Bernoulli/Binomial [2 possible outcomes]

n: number of trials p: probability of success x: number of successes

-Trials are independent.  
-The number of trials, n, is fixed.  
-Each trial outcome can be classified as a success or failure.  
-The probability of a success, p, is the same for each trial.

Expected value: np Variance: np(1-p) sd: sqrt(np(1-p)) Standard Error: sqrt[p(1-p)/n]

binomial mass function: (n x) p^x (1-p)^(n-x) where (n x) = n! / x!(n-x)!

(n r) "n choose r" in r code is choose (n,x)

choose(10,5)

## [1] 252

Gallup polls shows only 13% of employees are engaged at work. Among a random sample of 10 employees, what is the probability that 8 of them are engaged at work?

dbinom(8, size = 10, p = 0.13)

## [1] 2.778e-06

Binomial distribution can be approximated by normal distribution for large n (number of trials). A binomial distribution with at least 10 expected successes and 10 expected failures closely follows a normal distribution. I.e np >= 10 and n(1-p) >= 10

s = sqrt(np(1−p))

Example

n=160  
p=0.28  
mean=n\*p  
s=sqrt(p\*(1-p)\*(n))  
sum(dnorm(50:160, mean = mean, sd = s))

## [1] 0.2037

Example: Average FaceBook user has 245 friends. Probablility of a superuser friend (those who contribute more than average) is 0.25%. What is probability that a FaceBook user has 70 or more superuser friends?

Using the binomial distribution:

sum(dbinom(70:245, size = 245, p = 0.25))

## [1] 0.1128

Using the normal distribution:

sum(dnorm(70:245, mean = 61.25, sd = 6.78))

## [1] 0.1116

If probability of boy or girl birth is 50%, what's the probability of getting 7 or more girls out of 8 births?

pbinom(6, size = 8, prob = 0.5, lower.tail = FALSE)

## [1] 0.03516

### Normal Distribution

-Distribution is unimodal and symmetric  
-Points are on a straight line on a normal probability plot

Expected value: sum[x\* p(x)] Variance: sum[x^2 \* p(x)] - Expected Value^2

under the bell curve: 68% is within 1 standard deviation of the mean 95% is within 2 standard deviations of the mean 99.7% is within 3 standard deviations of the mean

90% = 1.28 \* sd + mean 95% = 1.645 \* sd + mean 97.5% = 1.96 \* sd + mean 99.7% = 2.33 \* sd + mean

**pnorm** returns percentile for an input observation/data point, x **qnorm** returns the observation for an input percentile

What is the critical value of a 95% CI given normal distribution?

qnorm(0.95, 0, 1)

## [1] 1.645

qnorm(0.95, 10, 1)

## [1] 11.64

If normal distribution with mean = 1020 and sd = 50, what is probability of getting more than 1160? Lower.tail=False means get the probability distribution curve to the right of that observation (otherwise pnorm returns probability distribution curve to the left of that observation)

pnorm(1160, mean = 1020, sd = 50, lower.tail = FALSE)

## [1] 0.002555

Equivalent:

1 - pnorm(50, mean = 45, sd = 3.2)

## [1] 0.05909

pnorm(50, mean = 45, sd = 3.2, lower.tail=FALSE)

## [1] 0.05909

Normal probability plot: the closer the points are to a perfect straight line, the more confident we can be that the data follow the normal model. -Left skewed - points bend down and to right of line -Right skewed - points bend up and to the left of the line -Short tails((narrower than the normal distribution) - points follow an S shaped-curve -Long tails (wider than the normal distribution) - points start below the line, bend to follow it, and end above it

## Foundations for Inference

sample statistics ---> point estimate ----> population parameter

*Central Limit theorem*: The distribution of sample statistics is nearly normal, centered at the population mean, and with a standard deviation equal to the population standard deviation divided by square root of the sample size

Independence: Sampled observations must be independent. -random sample/assignment -if sampling without replacement, n < 10% of population

Sample size/skew: Either the population distribution is normal, or if the population distribution is skewed, the sample size is large (rule of thumb: n > 30)

Confidence Interval: % Confident that the interval contains the unknown population mean.

Equals sample mean +/- margin of error (middle xx% of the normal distribution \* standard error of the sampling distribution)

Mean +/- z \* (s / sqrt(n)) ----> z \* (s / sqrt(n)) is called the margin of error where z is 1.96 for 95% CI (e.g.)

To determine the needed sample size to attain a given margin of error: n = ((s \* z) / ME)^2

qnorm(0.025)

## [1] -1.96

Remember that a 95% confidence interval means there is 95% under the normal curve with 5% split between the upper and lower tails (due to symmetry).

Avg temp in LA is 77 degrees in June with sd = 5. How cold are the coldest 20% of the days during June in LA?

qnorm(0.20, mean = 77, sd = 5)

## [1] 72.79

### Standard Normal (mean = 0 and sd = 1)

If X∼N(μ,σ2), then Z= (X−μ) / σ is standard normal (or (observation - mean) / sd). Used to compare results from different distributions.

### Bayes Theorem

P(A|B) = P(A and B). P(B)

### T Confidence Intervals

t quantiles are based on % confidence and degrees of freedom(df)

dept.stat.lsa.umich.edu/~kshedden/Courses/Stat401/Notes/t-quantiles.pdf

CI: mean +/- t quantile (for confidence and df=n-1) \* sd / sqrt(n)

Example: In a population of interest, a sample of 9 men yielded a sample average brain volume of 1,100cc and a standard deviation of 30cc. What is a 95% Student's T confidence interval for the mean brain volume in this new population?

1100 + c(-1, 1) \* qnorm(0.975) \* 30/3

## [1] 1080 1120

Remember that a 95% confidence interval means there is 95% under the normal curve with 5% split between the upper and lower tails (due to symmetry).

As CI increases, width of the interval increases, accuracy increases but precision decreases. Higher precision and higher accuracy through increased sample size.

It is much more common for a researcher to be interested in the difference between means than in the specific values of the means themselves.

A confidence interval on the difference between means is computed using the following formula:

Variance: S2p={(nx−1)S2x+(ny−1)S2y}/(nx+ny−2)

CI: [Mean1 - Mean2] +/- t quantile (on confidence and df n1 + n2 - 2) \* Sp \* (1/nx + 1/ny)^0.5

Example 1 - Unequal variance

n1 <- 100; n2 <- 100  
var1 <- .25  
mean1 <- 4  
var2 <- 4  
mean2 <- 6  
#pooled variance  
sp <- sqrt( ((n1 - 1) \* var1 + (n2-1) \* var2) / (n1 + n2-2))  
md <- mean1 - mean2  
semd <- sp \* sqrt(1 / n1 + 1/n2)  
  
md + c(-1, 1) \* qt(.975, n1 + n2 - 2) \* semd

## [1] -2.407 -1.593

Example 2 - Assume equal variance

n1 <- 9; n2 <- 9  
sd1 <- 1.5  
mean1 <- -3  
sd2 <- 1.8  
mean2 <- 1  
#average of two variances  
sp <- sqrt((1.5^2 + 1.8^2)/2)  
md <- mean1 - mean2  
semd <- sp \* sqrt(1 / n1 + 1/n2)  
  
md + c(-1, 1) \* qt(.975, n1 + n2 - 2) \* semd

## [1] -5.656 -2.344

### Hypothesis Testing

1. Set the null hypothesis ("nothing is going on) mu = null value and alternate hypothesis mu is >, < or not equal to null value
2. Calculate the point estimate (e.g. mean from sample)
3. Check conditions a. independence - random sample or assignment; n < 10% of population b. sample size/skew - n >= 30, larger if population very skewed
4. Draw sampling distribution, shade p-value, calculate test statistic z = (mu - mean) / (sd/sqrt(n))
5. Make decision and interpret in relation to research question a. if p-value < alpha, reject Ho b. if p-value > alpha, fail to reject Ho

Type 1 error: Rejecting Ho when it is true.  
Type 2 error: Failing to reject Ho when Ha is true.

Significance level (alpha) equals the probability of a type 1 error.

Probability of (Type 1 error | H0 true) = alpha.

Probability of type 2 error is β

Increasing alpha increases the Type 1 error rate but decreases the Type 2 error rate.

Power of a test is the probability of correctly rejecting H0 (because it is false), and the probability of doing so is 1 − β

Power increases as: -MuA gets further from Mu0 -n increases -alpha gets larger -one-sided (vs. two-sided) tests - this is because alpha is divided by 2

Effect size = (difference between means) / standard deviation

power.t.test(n=16, delta = 2, sd = 4, type = "one.sample", alt = "one.sided")

##   
## One-sample t test power calculation   
##   
## n = 16  
## delta = 2  
## sd = 4  
## sig.level = 0.05  
## power = 0.604  
## alternative = one.sided

where delta is difference between MuA and Mu0

A two-sides hypothesis test with threshhold of alpha is equivqlent to a CI with confidence level of 1 - alpha

A one-sided hypothesis test with threshhold of alpha is equivqlent to a CI with confidence level of 1 - (2 \* alpha)

Hypothesis testing for nearly normal point estimate:

Z = (point estimate - null value) / standard error

Real differences between the point estimate and null value are easier to detect with larger samples

### P-values

Most common measure of "statistical significance"

The P-value is the probability under the null hypothesis of obtaining evidence as extreme or more extreme than would be observed by chance alone.

Calculate the test statistic (z = mean1 - mean2 / standard error, for a normal distribution) and find prob of being above that test statistic (or above that statistic and below the negative of the statistic for a two-sided test). Find Z on the outer perimeter of the z-table, the prob of being less than that value is shown in the table. 1 - the table amount is the prob of being larger than the z-value.

If the P-value is small, then either H0 is true and we have observed a rare event or H0 is false

What is probability of getting a T statistic as large as 0.8 (df = 15)?

pt(0.8, 15, lower.tail = FALSE)

## [1] 0.2181

If the P-value is less than α, you reject the null hypothesis

For two sided hypothesis test, double the smaller of the two one sided hypothesis test P-values

If each gender has an independent 50% probability for each birth, what's the probability of getting 7 or more girls out of 8 births?

#Prob of more than 6 girls out of 8 births given prob of girl = 0.5  
pbinom(6, size = 8, prob = 0.5, lower.tail = FALSE)

## [1] 0.03516

Suppose that a hospital has an infection rate of 10 infections per 100 person/days at risk (rate of 0.1) during the last monitoring period. Assume that an infection rate of 0.05 is an important benchmark.

Given the model, could the observed rate being larger than 0.05 be attributed to chance?

Under H0:λ=0.05 so that λ0100=5 Consider Ha:λ>0.05.

#Prob of more than 9 infections given rate of 5 infections for 100 person/days.  
ppois(9, 5, lower.tail = FALSE)

## [1] 0.03183

### Resampling/Bootstrapping

Bootstrap procedure for calculating confidence interval for the median from a data set of observations

1. Sample observations **with replacement** from the observed data resulting in one simulated complete data set
2. Take the median of the simulated data set
3. Repeat these two steps times, resulting in simulated medians
4. These medians are approximately drawn from the sampling distribution of the median of observations; therefore we can
   * Draw a histogram of them
   * Calculate their standard deviation to estimate the standard error of the median
   * Take the and percentiles as a confidence interval for the median

Taking samples from existing data (n observations for B bootstrap resamples) with replacement

library(UsingR)

## Warning: package 'UsingR' was built under R version 3.1.1

## Loading required package: MASS  
## Loading required package: HistData  
## Loading required package: Hmisc  
## Loading required package: grid  
## Loading required package: lattice  
## Loading required package: survival  
## Loading required package: splines  
## Loading required package: Formula  
##   
## Attaching package: 'Hmisc'  
##   
## The following objects are masked from 'package:base':  
##   
## format.pval, round.POSIXt, trunc.POSIXt, units  
##   
## Loading required package: quantreg  
## Loading required package: SparseM

## Warning: package 'SparseM' was built under R version 3.1.1

##   
## Attaching package: 'SparseM'  
##   
## The following object is masked from 'package:base':  
##   
## backsolve  
##   
##   
## Attaching package: 'quantreg'  
##   
## The following object is masked from 'package:Hmisc':  
##   
## latex  
##   
## The following object is masked from 'package:survival':  
##   
## untangle.specials  
##   
##   
## Attaching package: 'UsingR'  
##   
## The following object is masked from 'package:survival':  
##   
## cancer

data(father.son)  
x <- father.son$sheight  
n <- length(x)  
B <- 10000  
#matrix of B resamples of size n (draw n \* B times) - B rows and n columns  
resamples <- matrix(sample(x, n \* B, replace = TRUE), B, n)  
#find median of each row and create new vector  
resampledMedians <- apply(resamples, 1, median)  
  
#Standard deviation and quantiles of the resampled medians  
sd(resampledMedians)

## [1] 0.08513

quantile(resampledMedians, c(.025, .975))

## 2.5% 97.5%   
## 68.43 68.82

## Histogram of bootstrap resamples

library(ggplot2)

##   
## Attaching package: 'ggplot2'  
##   
## The following object is masked from 'package:UsingR':  
##   
## movies

g = ggplot(data.frame(medians = resampledMedians), aes(x = resampledMedians))  
g = g + geom\_histogram(color = "black", fill = "lightblue", binwidth = 0.05)  
g

### A plot of the histrogram of the medians

# Inference

Load inference function from datacamp class. Inputs are: inference(variable, type = "ci", method = "simulation" or "theorectical", conflevel = 0.95, est = "mean" or "median, boot\_method = "se" or "perc")

inference(y = numeric variable, x = categorical variable, est = c("mean", "proportion"), type = "ht", null = 0, alternative = c("twosided","less", "greater"), method = c("simulation","theoretical"))

For simulation, inference needs success="yes" parameter (where "yes" is the label used in the data to represent a success. Default is 10,000 simulations.

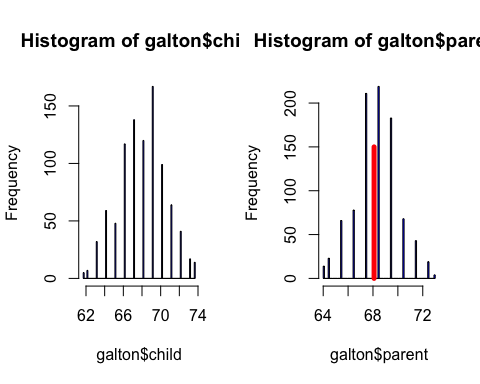
source("http://bit.ly/dasi\_inference")

## R Code

Subset and its count: nrow(subset(us12, us12$response == "atheist" ))

# Regression

library(UsingR)  
data(galton)  
par(mfrow=c(1,2))  
hist(galton$child, col='blue', breaks=100)  
hist(galton$parent,col='blue', breaks=100)  
  
meanChild <- mean(galton$child)  
lines(rep(meanChild, 100), seq(0,150,length=100),col='red', lwd=5)



meanParent <- mean(galton$parent)  
  
#Regression line through data  
fit <- lm(child ~ parent, data=galton)  
summary(fit)

##   
## Call:  
## lm(formula = child ~ parent, data = galton)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.805 -1.366 0.049 1.634 5.926   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 23.9415 2.8109 8.52 <2e-16 \*\*\*  
## parent 0.6463 0.0411 15.71 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.24 on 926 degrees of freedom  
## Multiple R-squared: 0.21, Adjusted R-squared: 0.21   
## F-statistic: 247 on 1 and 926 DF, p-value: <2e-16

#Check that mean of residuals is close to 0  
mean(fit$residuals)

## [1] -2.36e-15

#Check that residuals and predictor are uncorrelated   
cov(fit$residuals, galton$parent)

## [1] -1.79e-13

#Intercept  
intercept <- fit$coef[1]  
  
#Slope  
slope <- fit$coef[2]  
  
#Regression line  
fit <- lm(I(child - mean(child)) ~ I(parent - mean(parent)) - 1, data=galton)  
  
summary(fit)

##   
## Call:  
## lm(formula = I(child - mean(child)) ~ I(parent - mean(parent)) -   
## 1, data = galton)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.805 -1.366 0.049 1.634 5.926   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## I(parent - mean(parent)) 0.6463 0.0411 15.7 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.24 on 927 degrees of freedom  
## Multiple R-squared: 0.21, Adjusted R-squared: 0.21   
## F-statistic: 247 on 1 and 927 DF, p-value: <2e-16

Residuals, have mean zero. In other words, the residuals are "balanced" among the data points; they're just as likely to be positive as negative.

Residuals must be uncorrelated with our predictors

var(data)=var(estimate)+var(residuals)

varEst <- var(est(slope,intercept)) varRes <- var(fitchild) #check that variance of data equals variance of estimate + variance of residuals all.equal(varChild, varRes+varEst)

Linear regression Esimtate of child's height = β0 + Parent's Height β1

Yi is the observed value for child's height.

Use least squares to minimize the differences ∑i=1n[Yi−(β0+β1Xi)]^2

β1 = Cor(Y,X) / Sd(Y)Sd(X) #numerator is predictor variable β0 = Y − β1 \* X

Cor(X, Y) = Cor(Y, X) measures strenght of linear relationship between X and Y, stronger relationships closer to 1 and -1.

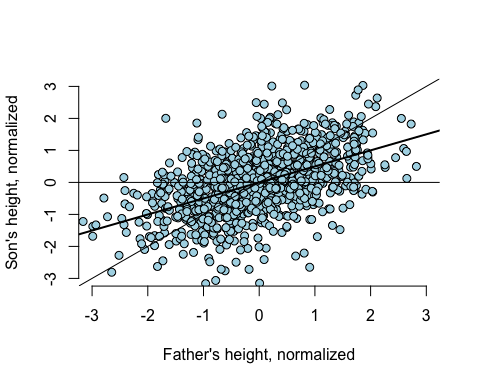
-1 <= Cor(X, Y) <= 1

y <- galton$child  
x <- galton$parent  
beta1 <- cor(y, x) \* sd(y) / sd(x)  
beta0 <- mean(y) - beta1 \* mean(x)  
rbind(c(beta0, beta1), coef(lm(y ~ x)))

## (Intercept) x  
## [1,] 23.94 0.6463  
## [2,] 23.94 0.6463

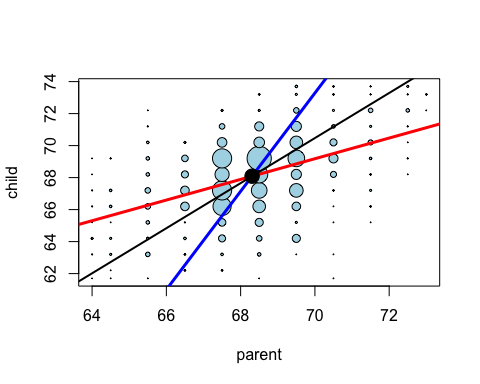
Regression to the mean

library(UsingR)  
#use father/son data only instead of parent/child  
data(father.son)  
#normalize the data  
y <- (father.son$sheight - mean(father.son$sheight)) / sd(father.son$sheight)  
x <- (father.son$fheight - mean(father.son$fheight)) / sd(father.son$fheight)  
#slope (correlation) does not need ratio of sds since normalized  
rho <- cor(x, y)  
myPlot <- function(x, y) {  
 plot(x, y,   
 xlab = "Father's height, normalized",  
 ylab = "Son's height, normalized",  
 xlim = c(-3, 3), ylim = c(-3, 3),  
 bg = "lightblue", col = "black", cex = 1.1, pch = 21,   
 frame = FALSE)  
}  
  
myPlot(x, y)  
abline(0, 1) # if there were perfect correlation (father's height 100% predicts son's height)  
abline(0, rho, lwd = 2) # father predicts son (shows regression toward the mean)  
abline(h = 0) # reference line for no relathionship (correlation is 0 so it is just a plot of a line at the mean)



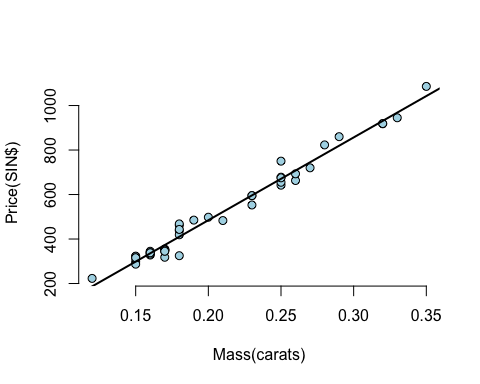
Unnormalied data

#plot the original Galton data points with larger dots for more freq pts  
y <- galton$child  
x <- galton$parent  
freqData <- as.data.frame(table(galton$child, galton$parent))  
names(freqData) <- c("child", "parent", "freq")  
plot(as.numeric(as.vector(freqData$parent)),   
 as.numeric(as.vector(freqData$child)),   
 pch = 21, col = "black", bg = "lightblue",  
 cex = .07 \* freqData$freq, xlab = "parent", ylab = "child")  
  
#original regression line, children as outcome, parents as predictor  
abline(mean(y) - mean(x) \* cor(y, x) \* sd(y) / sd(x), #intercept  
 sd(y) / sd(x) \* cor(y, x), #slope  
 lwd = 3, col = "red")  
  
#new regression line, parents as outcome, children as predictor  
abline(mean(y) - mean(x) \* sd(y) / sd(x) / cor(y, x), #intercept  
 sd(y) / cor(y, x) / sd(x), #slope  
 lwd = 3, col = "blue")  
  
#assume correlation is 1 so slope is ratio of std deviations  
abline(mean(y) - mean(x) \* sd(y) / sd(x), #intercept  
 sd(y) / sd(x), #slope  
 lwd = 2)  
points(mean(x), mean(y), cex = 2, pch = 19) #big point of intersection



The least squares estimate for μi = β0 + β1 \* xi is exactly the maximimum likelihood estimate (regardless of σ)

library(UsingR)  
data(diamond)   
plot(diamond$carat,diamond$price,  
 xlab="Mass(carats)",  
 ylab="Price(SIN$)",  
 bg="lightblue", col="black",cex=1.1,pch=21,frame=FALSE)  
abline(lm(price~carat,data=diamond),lwd=2)



fit<-lm(price~carat,data=diamond)   
coef(fit)

## (Intercept) carat   
## -259.6 3721.0

#strange intercept interpretation since there is no data on weight = 0  
#interpretation of B1 - for each 1 carat increase, an expected 3721.02 (SIN) dollar increase in price for every carat increase in mass of diamond  
  
fit2 <-lm (price~I(carat-mean(carat)),data=diamond)   
coef(fit2)

## (Intercept) I(carat - mean(carat))   
## 500.1 3721.0

#better interpretation of B0 after centering the data; $500.1 is the expected price for the average sized diamond of the data (0.2042 carats)  
  
#predicting the price of a diamond  
newx<-c(0.16,0.27,0.34) # three carats to price  
coef(fit)[1]+coef(fit)[2]\*newx #use regression equation on these carats

## [1] 335.7 745.1 1005.5

#OR  
  
#use predict function in R; "carat in function tells R which variable newx represents  
predict(fit,newdata=data.frame(carat=newx))

## 1 2 3   
## 335.7 745.1 1005.5

Residual, the differnce between the observed and predicted outcome; the vertical distance between the observed data point and the regression line

* Expected value of the residuals is 0
* Positive residuals are above the line, negative residuals are below
* Residual plots highlight poor model fit

data(diamond)   
y<-diamond$price  
x<-diamond$carat  
n<-length(y)   
fit<-lm(y~x)  
e<-resid(fit)  
yhat<-predict(fit)  
max(abs(e-(y-yhat))) #largest absolute difference between residual and difference between observed and fitted values

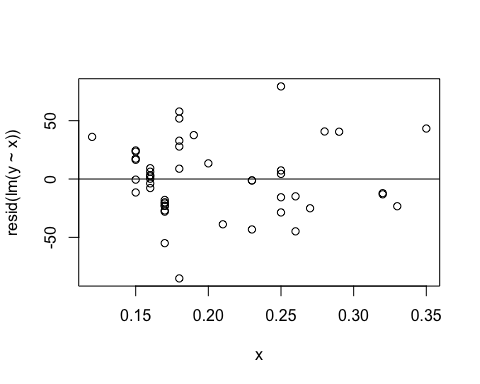
## [1] 9.486e-13

#OR plug into linear equation - same outcome  
max(abs(e-(y-coef(fit)[1]-coef(fit)[2]\*x)))

## [1] 9.486e-13

Residual plot

plot(x, resid(lm(y~x)))  
abline(h=0)



Estimated variance equals 1/n-2 times sum of residuals squared:

σ^2 = 1/n-2 ∑e^2

y<-diamond$price  
x<-diamond$carat  
n<-length(y)   
fit<-lm(y~x)  
summary(fit)$sigma

## [1] 31.84

#OR  
sqrt(sum(resid(fit)^2)/(n-2))

## [1] 31.84

R2 (R squared) The percentage of variation explained by the regression model

* R2 is the sample correlation squared
* 0 ≤ R2 ≤ 1
* R2 can be a misleading summary of model fit

Inference in Regression

library(UsingR); data(diamond)  
y <- diamond$price; x <- diamond$carat; n <- length(y)  
beta1 <- cor(y, x) \* sd(y) / sd(x)  
beta0 <- mean(y) - beta1 \* mean(x)  
e <- y - beta0 - beta1 \* x  
  
sigma <- sqrt(sum(e^2) / (n-2))   
ssx <- sum((x - mean(x))^2)  
  
seBeta0 <- (1 / n + mean(x) ^ 2 / ssx) ^ .5 \* sigma #standard error for B0  
seBeta1 <- sigma / sqrt(ssx) #standard error for B1  
  
tBeta0 <- beta0 / seBeta0; tBeta1 <- beta1 / seBeta1 #t statistics  
pBeta0 <- 2 \* pt(abs(tBeta0), df = n - 2, lower.tail = FALSE) #p-value for B0  
pBeta1 <- 2 \* pt(abs(tBeta1), df = n - 2, lower.tail = FALSE) #p-value for B1  
  
fit <- lm(y ~ x);   
summary(fit)$coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -259.6 17.32 -14.99 2.523e-19  
## x 3721.0 81.79 45.50 6.751e-40

## Getting a confidence interval

sumCoef <- summary(fit)$coefficients  
sumCoef[1,1] + c(-1, 1) \* qt(.975, df = fit$df) \* sumCoef[1, 2]

## [1] -294.5 -224.8

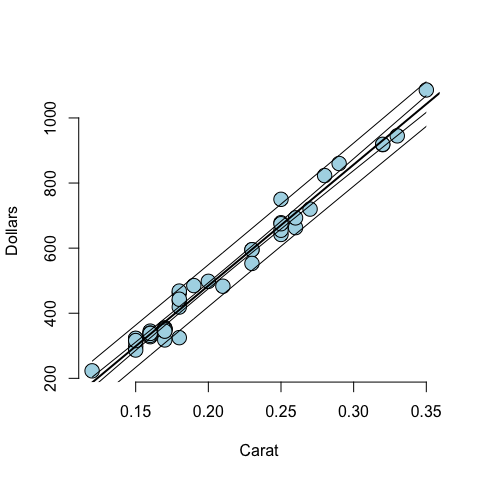
sumCoef[2,1] + c(-1, 1) \* qt(.975, df = fit$df) \* sumCoef[2, 2]

## [1] 3556 3886

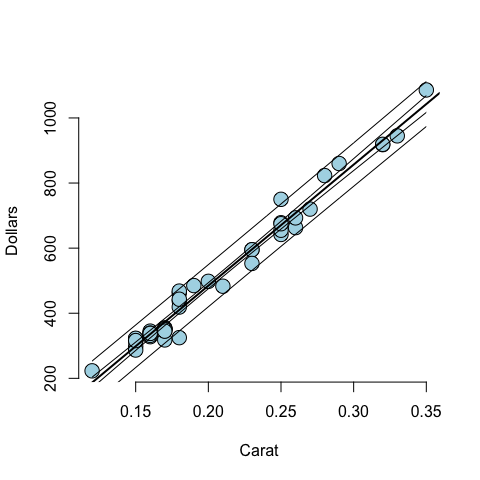
With 95% confidence, we estimate that a 0.1 carat increase in diamond size results in a 355.6 to 388.6 increase in price in (Singapore) dollars.

## Prediction Interval

The prediction interval must incorporate the variabilibity in the data around the line



Or using R function "predict"



# Multivariable regression

The interpretation of a multivariate regression coefficient is the expected change in the response per unit change in the regressor, holding all of the other regressors fixed

The least squares estimate for the coefficient of a multivariate regression model is exactly regression through the origin with the linear relationships with the **other regressors removed from both the regressor and outcome by taking residuals**

### Linear model with two variables and an intercept

#manually  
n <- 100; x <- rnorm(n); x2 <- rnorm(n); x3 <- rnorm(n)  
y <- x + x2 + x3 + rnorm(n, sd = .1)  
e <- function(a, b) a - sum( a \* b ) / sum( b ^ 2) \* b  
ey <- e(e(y, x2), e(x3, x2))  
ex <- e(e(x, x2), e(x3, x2))  
sum(ey \* ex) / sum(ex ^ 2)

## [1] 0.9913

#Using R functionality  
ey <- resid(lm(y ~ x2 + x3 - 1))  
ex <- resid(lm(x ~ x2 + x3 - 1))  
sum(ey \* ex) / sum(ex ^ 2)

## [1] 0.9913

#OR  
  
coef(lm(y ~ x + x2 + x3 - 1)) #the -1 removes the intercept term

## x x2 x3   
## 0.9913 0.9936 1.0196

So that the interpretation of a multivariate regression coefficient is the expected change in the response per unit change in the regressor, holding all of the other regressors fixed