Power of a test and Regression

In this PBL we will explore the concept of power of an experiment, and examine regression in great detail.

Goals of PBL **BIOL2006, PBL 2**

Sample size

We will learn a simple way to calculate desired sample sizes with a given power

Regression 1

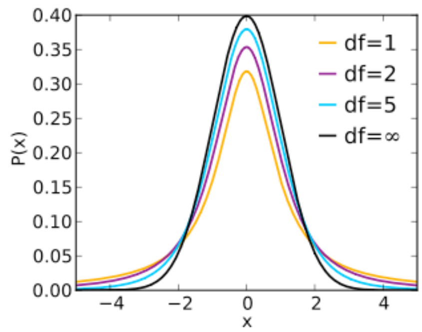
We will discuss single linear models using a continuous variable

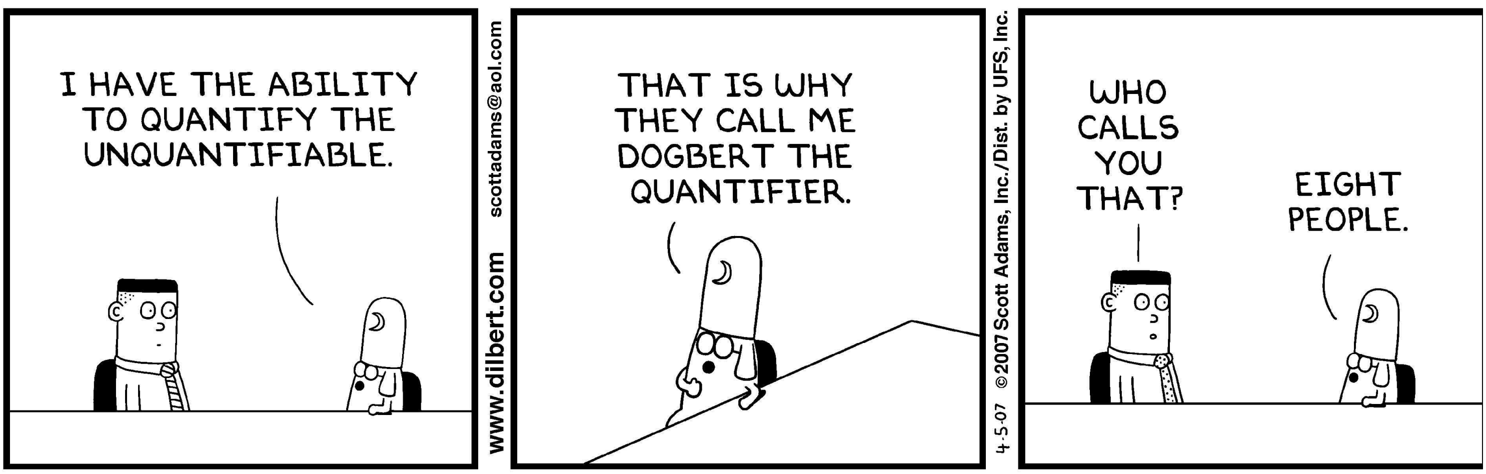
Regression 2

We will discuss more complex linear models using a two nominal variables

Regression 3

We will discuss the hypothesis testing framework for regression





Sample size, errors, and power

Activity I

In many circumstances your experimental design will require great efforts from you and collaborators. As such, it is a great idea if you design your experiments with the confidence that you have enough power to reject the null hypothesis. However, it is difficult to know in advance how to calculate power, as this usually requires some previous information about the expected difference between estimated parameters, and perhaps an indication of whether the difference between means goes in certain direction. Pilot experiments are great for this!

Take a look at the figure and its explanation on the next two pages. This is called a Nomogram, and it will help you understand several important concepts including type I, type II error, and power.

To use the Nomogram you will need to:

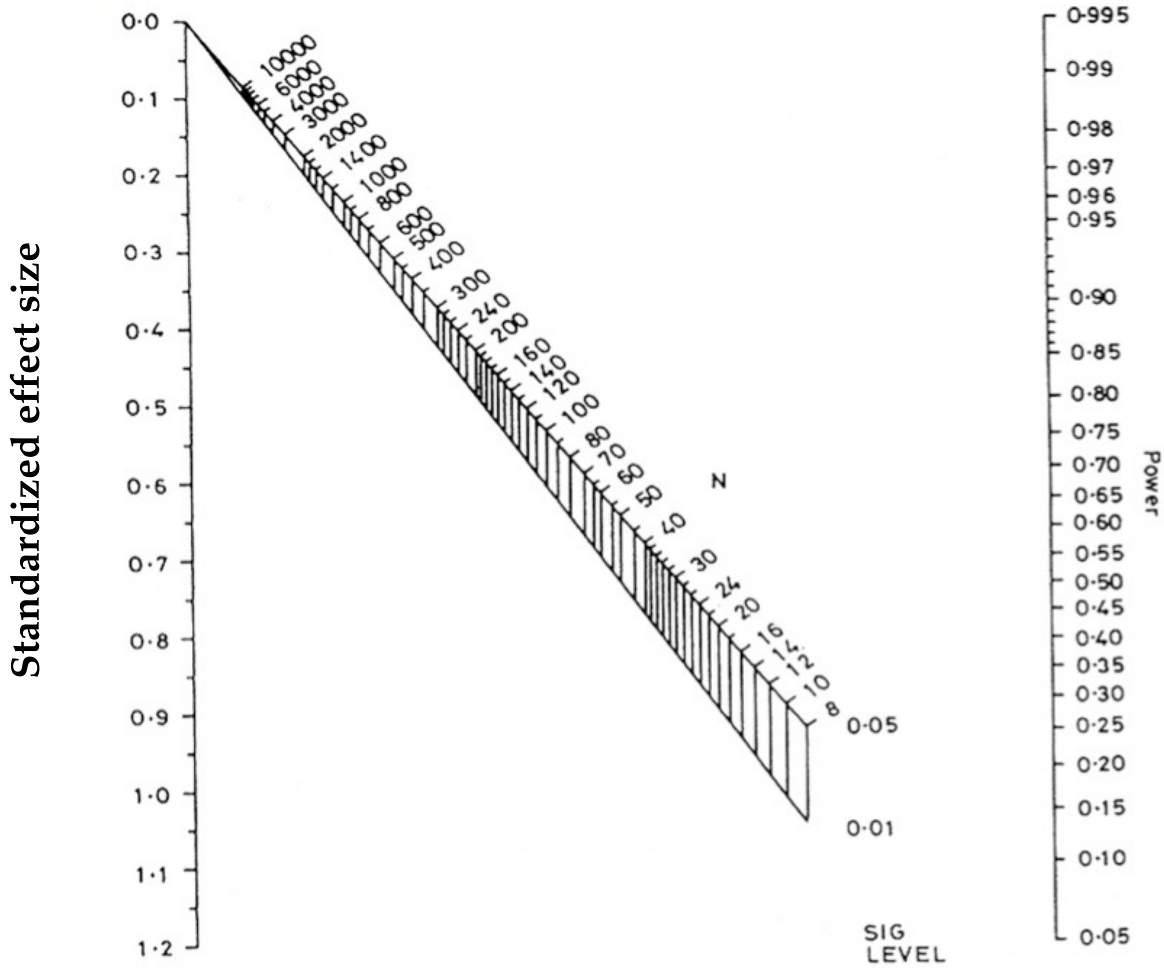
1) Define the effect size of your experiment
2) Define the required power for your experiment - often 80%
3) Define the significance level for rejecting your null hypothesis
4) Now, trace a line (use a ruler) between the desired effect size and the expected power of your experiment. Note where this line intersects the diagonal in the middle of the page. Make sure you pay attention to the line corresponding to the significance level you chose for your hypothesis testing.

Questions for discussion:

1) How does sample size affect your type I error?
2) How does sample size affect your type II error?
3) Discuss some key elements that will maximise the power of your experiment
4) If you had limited resources (i.e., small N), could you still increase the power of your experiment?

Sample size, errors, and power

We will use a simple approach to calculate power and understand the concept behind type I and type II errors



Nomogram for calculating sample size or power. Reproduced from Altman [5], with permission.

Sample size, errors, and power

INFORMATION ABOUT THE FIGURE

The left Y-axis

This axis denotes the standardized effect size between the means of two groups

The right Y-axis

This axis denotes the power of the test

The upper diagonal

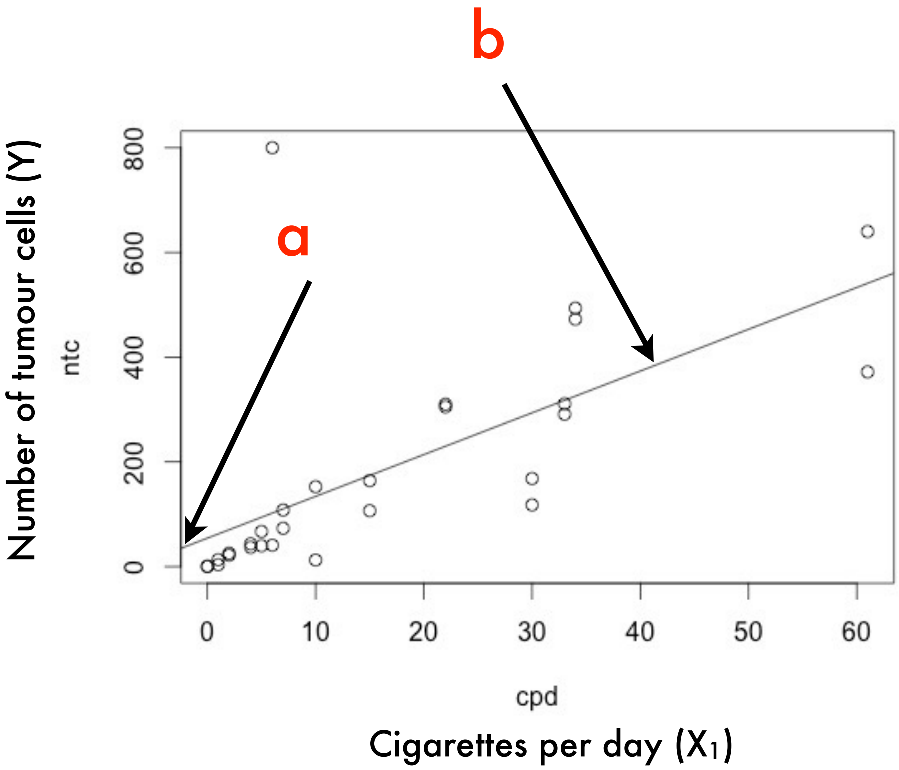
This diagonal represents the required sample sizes for a rejection probability of 0.05 and a given combination of effect size and power

The lower diagonal

This diagonal represents the required sample sizes for a rejection probability of 0.01 and a given combination of effect size and power

Regression

BIOL2006



Regression analyses and Linear Models

In our previous lecture we learned that two numeric variables, a predictor and a response can be related to one another.

The relationship between these variables is expressed best by a simple linear equation:

$$Y = bX_1 + a \quad \text{(eq1)}$$

Where b is the slope of the line, and a is the intercept. Note, that X and Y are given as they are your data points on which you are testing a hypothesis.

First POINT:

Equation (1) above is an example of what is known as Linear Models.

Linear Model (LM)

In a LM there can be more than one X variable. For example:

$Y = b_1X_1 + b_2X_2 + a$ (eq2)

If we look at the graph above, X_1 can be cigarettes per day, whereas X_2 (not shown) could be a different variable that we believe might also have an effect on cancer; for instance, age.

Second POINT:

The X variables in the LM can be either continuous, such as the variables in the graph above, or nominal, such as gender, geographic location, bench, research teams, etc.

<u>R and LMs</u>

R finds the Linear Model (LM) that best fits your data.

As such, R gives you estimates for b1, b2, and a, if you consider equation (2)

Activity 2: Linear Model, 1 continuous X

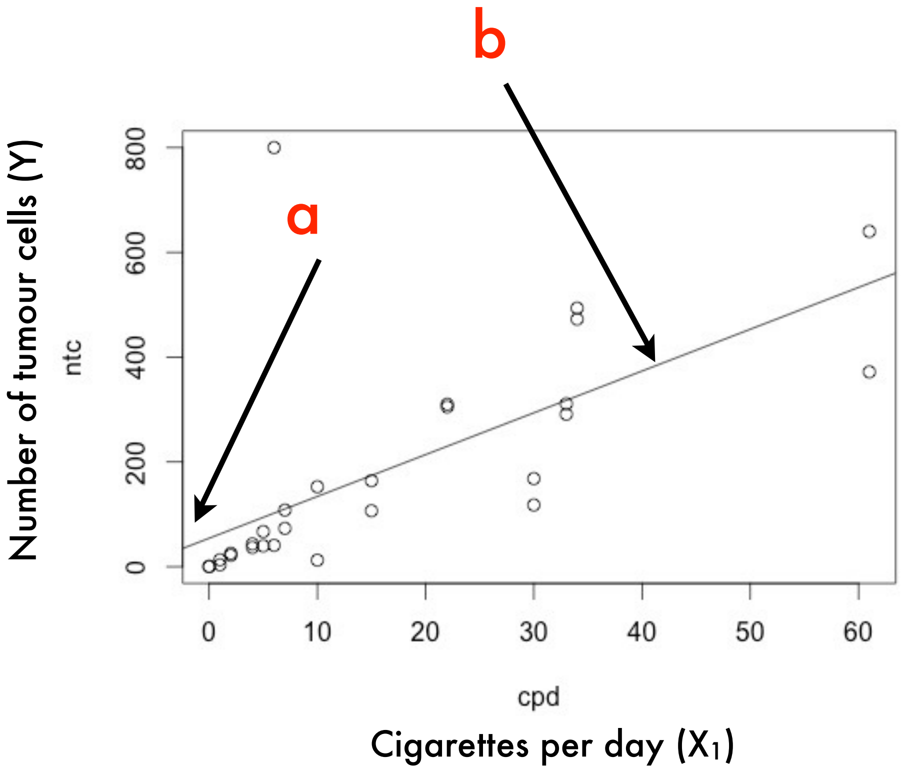
Notes: The R code on the next page is for illustrative purposes. Work with the output below, and notice that there is missing information. Work with your partner and if you have any questions, ask your tutor to give you some help.

Question 1: Write down the equation that describes the least square regression for the data in example 1 below.

Question 2: Why are there two t-tests in the table above? How are the t-values calculated? Calculate them and their associated *p-values*.

Question 3: Calculate the F value for the ANOVA table above and its associated *p-value*?

Question 4: What can you say about the relationship between smoking and cancer according to the results above? Please comment both on the null hypothesis being tested and also on the fit of the model (hint: Also calculate R^2).

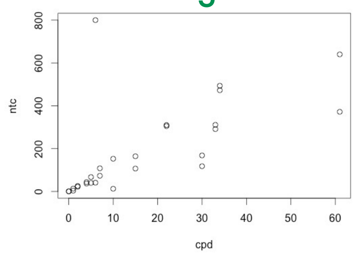


$$Y = b_1X_1 + a$$

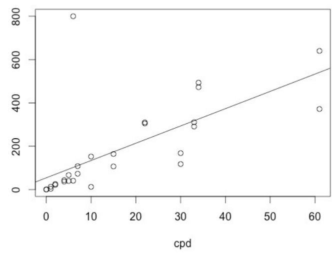
LM Exercise 1: 1 continuous X variable

$Y = b_1X_1 + a$ (eq3)

Scattergram



Best Fit




```
#Import datafile
#Make your data set default
attach (d3)
#Plot a scattergram of your variables
plot(ntc ~ cpd, data= d3) # Number of tumour cells by cigarettes per day
#Fit the best line to your scatterplot
abline(lm(ntc ~ cpd, data= d3)) #Remember to define your data set
#Find the regression parameters
fit <- lm(ntc~cpd) #Define the object fit with the regression function lm
summary (fit) #Obtain an estimate of the parameters for the LM
```

R output

```
Call:
lm(formula = ntc ~ cpd)

Residuals:
    Min      1Q  Median      3Q     Max
-175.82  -55.58  -43.73    3.03  697.76

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   54.382     42.659
cpd            7.977      1.807
---

R-squared:
F-statistic: 19.5 on 1 and 26 DF,p-value: 0.0001572
```

```
anova (fit)  #Obtain an ANOVA table - like the one we built during lecture
Analysis of Variance Table

Response: ntc
          Df Sum Sq Mean Sq F value    Pr(>F)
cpd        1 512522
Residuals 26 683510
```

Activity 2: Linear Model, 2 nominal X variables

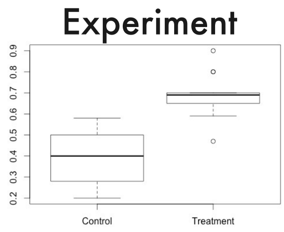
Question 1: Write down the equation that describes the least square regression for the data in example 2 on the next page.

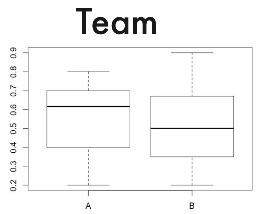
Question 2: What can you conclude about the effect of Team and Experiment on the Rate response?

Question 3:

1) Is there a significant interactions between TEAM and EXPERIMENT?
2) Calculate the mean for each group using the information above.
3) Draw in a single Bar Graph the relationship amongst the four groups, and add approximate standard errors to each mean.

Category	Equation	Mean	n	Std. Dev
Control Team A	Y_{00} =		7	0.141421356
Control Team B	Y_{01} =		6	0.123558353
Treatment Team A	Y_{10} =		7	0.058145958
Treatment Team B	Y_{11} =		6	0.14207979





Example 2: 2 nominal X variables X_1 and X_2

$$Y = b_1X_1 + b_2X_2 + b_3X_1X_2 + a \quad \text{(eq4)}$$

Interaction term

#Import data set, Example 2
#attach file
attach(example2)
#Plot the boxplot for Experiment vs. Rate
plot(Experiment,Rate)
#Plot the boxplot for Team vs. Rate
plot(Team,Rate)
#Fit the best LM to your data
fit <- lm(Rate~Experiment*Team, data=example2) #Define the object fit with the regression function lm
summary (fit) #Obtain an estimate of the parameters for the LM

R output

```
Call:
lm(formula = Rate ~ Experiment * Team, data = example2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.196667 -0.067262 -0.008333  0.070238  0.233333

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                0.38000    0.04541
ExperimentTreatment        0.33857    0.06422
TeamB                     -0.01667    0.06684
ExperimentTreatment:TeamB -0.03524    0.09453
---
R-squared: 0.6832
F-statistic: 15.82 on 3 and 22 DF,  p-value: 1.051e-05
```

anova (fit) #Obtain an ANOVA table - like the one we built during lecture

```
Analysis of Variance Table
     Response: Rate
                Df  Sum Sq Mean Sq F value    Pr(>F)
Experiment       1 0.67523 0.67523
Team             1 0.00760 0.00760
Experiment:Team  1 0.00201 0.00201
Residuals       22 0.31755 0.01443
```

Student's t table

PERCENTAGE POINTS OF THE T DISTRIBUTION

Tail Probabilities									
One Tail		0.10	0.05	0.025	0.01	0.005	0.001	0.0005	
Two Tails		0.20	0.10	0.05	0.02	0.01	0.002	0.001	
D	1	3.078	6.314	12.71	31.82	63.66	318.3	637	1
E	2	1.886	2.920	4.303	6.965	9.925	22.330	31.6	2
G	3	1.638	2.353	3.182	4.541	5.841	10.210	12.92	3
R	4	1.533	2.132	2.776	3.747	4.604	7.173	8.610	4
E	5	1.476	2.015	2.571	3.365	4.032	5.893	6.869	5
E	6	1.440	1.943	2.447	3.143	3.707	5.208	5.959	6
S	7	1.415	1.895	2.365	2.998	3.499	4.785	5.408	7
	8	1.397	1.860	2.306	2.896	3.355	4.501	5.041	8
O	9	1.383	1.833	2.262	2.821	3.250	4.297	4.781	9
F	10	1.372	1.812	2.228	2.764	3.169	4.144	4.587	10
	11	1.363	1.796	2.201	2.718	3.106	4.025	4.437	11
F	12	1.356	1.782	2.179	2.681	3.055	3.930	4.318	12
R	13	1.350	1.771	2.160	2.650	3.012	3.852	4.221	13
E	14	1.345	1.761	2.145	2.624	2.977	3.787	4.140	14
E	15	1.341	1.753	2.131	2.602	2.947	3.733	4.073	15
D	16	1.337	1.746	2.120	2.583	2.921	3.686	4.015	16
O	17	1.333	1.740	2.110	2.567	2.898	3.646	3.965	17
M	18	1.330	1.734	2.101	2.552	2.878	3.610	3.922	18
	19	1.328	1.729	2.093	2.539	2.861	3.579	3.883	19
	20	1.325	1.725	2.086	2.528	2.845	3.552	3.850	20
	21	1.323	1.721	2.080	2.518	2.831	3.527	3.819	21
	22	1.321	1.717	2.074	2.508	2.819	3.505	3.792	22
	23	1.319	1.714	2.069	2.500	2.807	3.485	3.768	23
	24	1.318	1.711	2.064	2.492	2.797	3.467	3.745	24
	25	1.316	1.708	2.060	2.485	2.787	3.450	3.725	25
	26	1.315	1.706	2.056	2.479	2.779	3.435	3.707	26
	27	1.314	1.703	2.052	2.473	2.771	3.421	3.690	27
	28	1.313	1.701	2.048	2.467	2.763	3.408	3.674	28
	29	1.311	1.699	2.045	2.462	2.756	3.396	3.659	29
	30	1.310	1.697	2.042	2.457	2.750	3.385	3.646	30
	32	1.309	1.694	2.037	2.449	2.738	3.365	3.622	32
	34	1.307	1.691	2.032	2.441	2.728	3.348	3.601	34
	36	1.306	1.688	2.028	2.434	2.719	3.333	3.582	36
	38	1.304	1.686	2.024	2.429	2.712	3.319	3.566	38
	40	1.303	1.684	2.021	2.423	2.704	3.307	3.551	40
	42	1.302	1.682	2.018	2.418	2.698	3.296	3.538	42
	44	1.301	1.680	2.015	2.414	2.692	3.286	3.526	44
	46	1.300	1.679	2.013	2.410	2.687	3.277	3.515	46
	48	1.299	1.677	2.011	2.407	2.682	3.269	3.505	48
	50	1.299	1.676	2.009	2.403	2.678	3.261	3.496	50
	55	1.297	1.673	2.004	2.396	2.668	3.245	3.476	55
	60	1.296	1.671	2.000	2.390	2.660	3.232	3.460	60
	65	1.295	1.669	1.997	2.385	2.654	3.220	3.447	65
	70	1.294	1.667	1.994	2.381	2.648	3.211	3.435	70
	80	1.292	1.664	1.990	2.374	2.639	3.195	3.416	80
	100	1.290	1.660	1.984	2.364	2.626	3.174	3.390	100
	150	1.287	1.655	1.976	2.351	2.609	3.145	3.357	150
	200	1.286	1.653	1.972	2.345	2.601	3.131	3.340	200
Two Tails		0.20	0.10	0.05	0.02	0.01	0.002	0.001	
One Tail		0.10	0.05	0.025	0.01	0.005	0.001	0.0005	
Tail Probabilities									