

# Data Analysis in the Health Sciences

Final Exam 2011 – EPIB-621

Student's Name: \_\_\_\_\_

Student's Number: \_\_\_\_\_

## INSTRUCTIONS

This examination consists of 8 questions on 19 pages, including this one. Tables of the normal distribution are provided on the last page. Please write your answers (NEATLY) in the spaces provided. Fully explain all of your answers. Each question is worth 10 points, for a total of 80.

1. \_\_\_\_\_
2. \_\_\_\_\_
3. \_\_\_\_\_
4. \_\_\_\_\_
5. \_\_\_\_\_
6. \_\_\_\_\_
7. \_\_\_\_\_
8. \_\_\_\_\_

Total (out of 80) \_\_\_\_\_

1. Suppose that the proportion of male births across northern populations has typically been 51%, compared to 49% females. It is hypothesized that this proportion may depend on the amount of a certain pollutant in the area. Data are collected from 100 communities in different regions of the north. For each community, the proportion of male births ( $y$ ) and the amount of the pollutant ( $x$ ) measured on a scale from 0 to 100 is collected. A change of 0.5% would be considered as clinically important. The following are the results from a linear regression model relating  $x$  to  $y$ :

```
> summary(x)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 11.92  25.68   46.20   46.52  63.21   88.94

Call:
lm(formula = y ~ x)

Residuals:
      Min       1Q   Median       3Q      Max
-0.0239212 -0.0068155  0.0004083  0.0074820  0.0206297

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.100e-01  2.456e-03  207.656  <2e-16 ***
x            -1.483e-06  4.760e-05  -0.031   0.975
```

(a) Report the intercept and slope from the model, each with 95% confidence interval.

(b) Overall, what would you conclude about the effect of the pollutant on the proportion of male births?

2. The following data were collected in 50 American states in 1964. The goal was to construct a linear model for the number of deaths in car accidents in order to better understand the effects of the various factors studied. Each observation corresponds to one state (i.e. there are 50 observations) and the list of variables is the following:

- Death = Number of deaths in car accidents (outcome variable)
- License (lic) = Number of people with a driving license in the state (in tens of thousands)
- Network (net)= Length of the road network (in thousands of miles)
- Gender (gend)= indicator variable indicating if there are more men than women in the state (value 1) or not (value 0)
- Gas = Gas consumption (in tens of millions gallons)

With 4 independent variables, there are a total of 15 possible models (without counting the model with only an intercept). The  $R^2$ , the adjusted  $R^2$ , BIC and the AIC values are provided in the table below.

Model	adj.r2	r2	aic	bic
Lic	0.9119	0.9137	703.40	709.14
Net	0.3500	0.3633	803.35	809.09
Gend	0.1049	0.1232	819.354	825.09
Gas	0.5514	0.5605	784.82	790.55
Lic+Net	0.9326	0.9354	690.96	698.61
Lic+Gend	0.9102	0.9139	705.31	712.95
Lic+Gas	0.9112	0.9148	704.79	712.44
Net+Gend	0.4336	0.4567	797.42	805.07
Net+Gas	0.6030	0.6192	779.65	787.30
Gend+Gas	0.5464	0.5649	786.32	793.97
Lic+Net+Gend	0.9322	0.9364	692.20	701.76
Lic+Net+Gas	0.9367	0.9406	688.75	698.31
Lic+Gend+Gas	0.9096	0.9152	706.57	716.13
Net+Gend+Gas	0.6067	0.6308	780.11	789.67
Lic+Net+Gend+Gas	0.9379	0.9430	688.72	700.19

(a) Using the BIC criterion, which model would be selected?

(b) Using the AIC criterion, which model would be selected?

(c) Using the adjusted R squared criterion, which model would be selected?

(d) Explain in your own words why the  $R^2$  value always increases when more variables are added to a model.

3. Consider a study examining the frequency of infections after c-section deliveries. Potential risk factors include:

- If the c-section was planned or not
- If the mother had risk factors such as diabetes or obesity.
- If preventive antibiotics were given prior to the c-section

The data are summarized in the following table:

		Planned c-section		Non planed c-section	
		Infection	No infection	Infection	No infection
Antibiotics	Mother at risk	1	17	11	87
Antibiotics	Mother not at risk	0	2	0	0
No antibiotics	Mother at risk	28	30	23	3
No antibiotics	Mother not at risk	8	32	0	9

A logistic regression model was fit to the data, and the R output is given below:

```
Call: glm(formula = infection ~ planned + antibio + risk, family = binomial)
```

Coefficients:

```
(Intercept)      planned      antibio      risk
   -0.8207      -1.0720     -3.2544      2.0299
```

```
Degrees of Freedom: 250 Total (i.e. Null); 247 Residual
```

```
Null Deviance:      299
```

```
Residual Deviance: 226.5      AIC: 234.5
```

(a) Interpret the estimated intercept (or a function of the intercept) in a way that is clinically meaningful.

(b) Calculate the odds ratios corresponding to each of the estimated coefficients for planned, antibio and risk. Provide a clinically meaningful interpretation of the point estimates of these odds ratios.

(c) A model with a planned\*risk interaction was fitted to the data. The R output is below:

```
Call: glm(formula = infection ~ planned + antibio + risk + planned * risk,
          family = binomial)
```

Coefficients:

(Intercept)	planned	antibio	risk	planned:risk
-16.566	15.178	-3.829	18.389	-17.027

Degrees of Freedom: 250 Total (i.e. Null); 246 Residual

Null Deviance: 299

Residual Deviance: 216.5           AIC: 226.5

Calculate the odds ratio corresponding to each of the four possible combinations of the risk/planned variables.



4. A group of 781 subjects were selected to participate in a study of myocardial infarction. The results are shown in the table below:

	Myocardial infarction	No myocardial infarction	Total
Smoker (or former smoker)	172	173	345
Never smoked	90	346	436
	262	519	781

Suppose that one wants to fit a logistic regression model to these data, using the model below:

$$\text{logit} \left\{ \frac{\pi(x)}{1 - \pi(x)} \right\} = \beta_0 + \beta_1 x,$$

where  $x$  represents the smoking variable (0/1) and  $\pi(x)$  represents the probability of a myocardial infarction for a given value of  $x$ .

(a) Using the data above, estimate the coefficient  $\beta_0$  of the logistic regression model.

(b) Using the data above, estimate the coefficient  $\beta_1$  of the logistic regression model.

5. The following data concern a sample of 337 subjects drawn from a cohort study of the incidence of coronary heart disease (CHD). This data set contains the following dependent and independent variables:

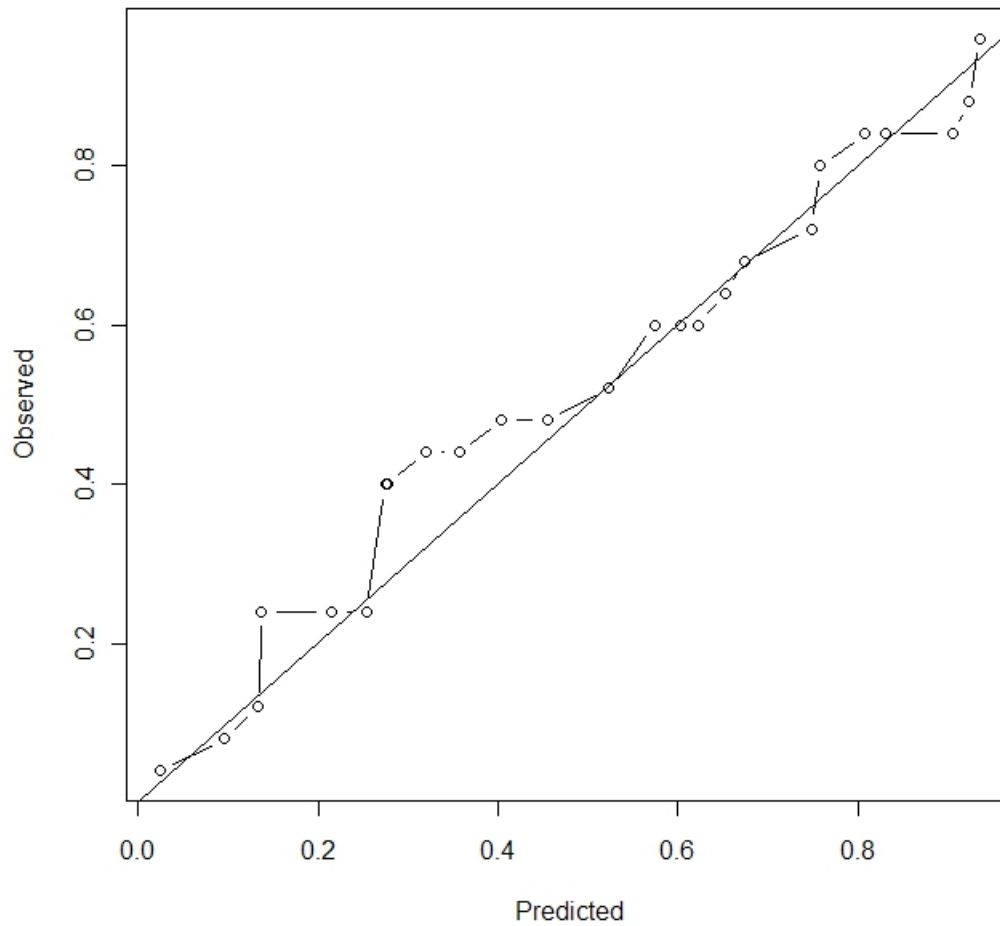
- years: number of years at risk
- energy: total energy intake (KCal per day/100)
- height: (cm)
- weight: (kg)
- fat: fat intake (g/day)
- fibre: dietary fibre intake (g/day)
- chd: CHD event (1=CHD event, 0=no event)

Seven years after baseline data collection, follow-up measurements were taken on members of the original sample that could be found and agreed to further participate. Out of the initial 337 subjects, 246 of them were followed up. The objective of this question is to investigate whether the 91 missing subjects are missing at random or not. In order to do so, a series of logistic regression models were fit using the indicator whether the subject was followed up or not as an outcome, and the variables year, energy, height, weight, fat, fibre and chd as explanatory variables. The table below shows the odds ratios and their confidence intervals for the 7 simple logistic models:

Variable	Odds Ratio	Confidence interval
years	1.07	(1.02,1.13)
energy	0.98	(0.93,1.03)
height	1.02	(0.98,1.06)
weight	1.00	(0.98,1.02)
fat	1.00	(0.90,1.11)
fibre	1.46	(0.90,2.35)
chd	0.33	(0.17,0.64)

In view of the results above, would you say that the data are missing completely at random? Explain your reasoning.

6. The plot below is a Hosmer-Lemeshow plot created by taking a study with sample size 625, and dividing it into 25 categories of 25 data points each.



(a) Overall, does the fit of the model seem reasonable?

(b) One point on the plot is  $(0.27, 0.4)$ . Do you think the discrepancy of 0.13 between observed and predicted for this point is within the bounds of chance? Provide a quantitative argument to back up your answer.

7. Past data suggest that a medication for secondary stroke prevention may have different effectiveness in different populations from different regions. A clinical trial is carried out on 2500 prior stroke patients, 500 subjects from each of 5 regions. All subjects are followed for one year to determine if they suffer a second stroke ( $y = 1$ ) or not ( $y = 0$ ), depending on whether they are taking the medication ( $x = 1$ ) or a placebo ( $x = 0$ ). The following hierarchical logistic regression model is run:

```

model
{
  for (j in 1:5)
  {
    for (i in index[j]:index2[j])
    {
      logit(p[i]) <- alpha[j] + beta[j]*x[i]
      y[i] ~ dbern(p[i])
    }
    alpha[j] ~ dnorm(mu.a, tau.a)
    beta[j] ~ dnorm(mu.b, tau.b)
  }
  mu.a ~ dnorm(0,0.001)
  tau.a <- 1/(sigma.a*sigma.a)
  sigma.a ~ dunif(0,20)
  mu.b ~ dnorm(0,0.001)
  tau.b <- 1/(sigma.b*sigma.b)
  sigma.b ~ dunif(0,20)

  alpha12 <- alpha[1] - alpha[2]
  alpha13 <- alpha[1] - alpha[3]
  alpha14 <- alpha[1] - alpha[4]
  alpha15 <- alpha[1] - alpha[5]
  alpha23 <- alpha[2] - alpha[3]
  alpha24 <- alpha[2] - alpha[4]
  alpha25 <- alpha[2] - alpha[5]
  alpha34 <- alpha[3] - alpha[4]
  alpha35 <- alpha[3] - alpha[5]
  alpha45 <- alpha[4] - alpha[5]
  beta12 <- beta[1] - beta[2]
  beta13 <- beta[1] - beta[3]
  beta14 <- beta[1] - beta[4]
  beta15 <- beta[1] - beta[5]
  beta23 <- beta[2] - beta[3]
  beta24 <- beta[2] - beta[4]
  beta25 <- beta[2] - beta[5]
  beta34 <- beta[3] - beta[4]

```

```

beta35 <- beta[3] - beta[5]
beta45 <- beta[4] - beta[5]
}

# Inits

list(alpha=c(0,0,0,0,0), beta=c(0,0,0,0,0), mu.a=0, sigma.a = 1,
mu.b=0, sigma.b = 1)

# Data

list(index = c(1, 501, 1001, 1501, 2001),
index2 =c(500, 1000, 1500, 2000, 2500),
x = c(0, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0,
0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1,
....etc...
0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1,
0, 0, 1, 1, 1, 1, 1, 1), y = c(0, 1, 0, 1, 0, 0, 1, 1, 0, 0,
0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0,
....etc...
0, 0, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1,
0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0))

```

## Results

node	mean	sd	2.5%	median	97.5%
alpha[1]	0.1581	0.08687	-0.03387	0.1634	0.3148
alpha[2]	0.2215	0.08688	0.06622	0.2158	0.4127
alpha[3]	0.2185	0.08432	0.06289	0.2144	0.3987
alpha[4]	0.2097	0.08507	0.05057	0.2057	0.3934
alpha[5]	0.1459	0.091	-0.06247	0.155	0.2979
alpha12	-0.0634	0.1078	-0.3349	-0.0344	0.09792
alpha13	-0.06038	0.1048	-0.3193	-0.0329	0.1057
alpha14	-0.05161	0.1066	-0.3174	-0.02361	0.1179
alpha15	0.01216	0.09255	-0.1737	0.00327	0.2287
alpha23	0.00301	0.09265	-0.1882	8.036E-5	0.2127
alpha24	0.01178	0.09249	-0.1826	0.00399	0.2189
alpha25	0.07556	0.117	-0.0857	0.04047	0.3724
alpha34	0.00877	0.09595	-0.2005	0.003673	0.2145
alpha35	0.07255	0.1101	-0.0865	0.04215	0.3514
alpha45	0.06378	0.115	-0.1065	0.03106	0.3591
beta[1]	0.1687	0.1207	-0.0671	0.1661	0.4116
beta[2]	0.117	0.1272	-0.1604	0.1248	0.3519
beta[3]	0.211	0.1238	-0.0155	0.2032	0.4744
beta[4]	0.07009	0.1384	-0.2366	0.08483	0.3103



beta[5]	0.2494	0.1413	0.0144	0.2348	0.5694
beta12	0.0517	0.1486	-0.2126	0.02476	0.4046
beta13	-0.04231	0.1424	-0.3595	-0.0243	0.2413
beta14	0.09859	0.1609	-0.1519	0.06066	0.4875
beta15	-0.08074	0.1512	-0.4419	-0.04739	0.1675
beta23	-0.09401	0.1519	-0.459	-0.06086	0.1498
beta24	0.0469	0.1433	-0.2235	0.02436	0.3748
beta25	-0.1324	0.1794	-0.5705	-0.08546	0.1212
beta34	0.1409	0.1705	-0.1004	0.1028	0.5482
beta35	-0.03842	0.1472	-0.386	-0.01499	0.2381
beta45	-0.1793	0.2006	-0.6683	-0.1299	0.07878
sigma.a	0.1043	0.1261	0.00255	0.07071	0.414
sigma.b	0.1715	0.1833	0.00565	0.1247	0.6222

(a) Overall, is there evidence for differences in stroke rates amongst the five regions? Explain why or why not.

(b) Overall, is there evidence for different effectivenesses of the medication between the five regions? Explain why or why not.

8. Two unrelated studies are being planned to estimate the effect of blood pressure on coronary heart disease (CHD), while adjusting for potential confounding effects from variables such as age and cholesterol. In one study, Researcher A will measure blood pressure one time within each of 100 randomly selected patients from his hypertension (high blood pressure) clinic. In the second study, Researcher B will include a random sample of 200 normotensive (normal blood pressure) subjects, measuring blood pressure at three different times, and using the average of these three measures as his main independent variable. Both studies will follow subjects for one year, recording the presence or absence of CHD up to that time.

Discuss the advantages and disadvantages of the two studies by comparing the likely consequences of the various differences in study designs.

## Normal Density Table

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Table of standard normal distribution probabilities. Each number in the table provides the probability that a standard normal random variable will be less than the number indicated.