Datasets and Instructions

Class Project

Directions

Choose **only one** of the questions and use the provided `RMardown Project Template` or a similar LaTeX Template to write your project report.

For the selected Data (question),

1. Create a title of your research question from the objective of the study.
2. The response and exposure variables for each dataset are provided. The identification number (`idnum`) variable is not part of the covariates of interest.
3. Fit appropriate statistical model(s). (See provided Hint/suggestions.) Explore the data. You may have to transform the response variable or covariates and/or standardize some covariates if necessary. Check for correlations among the variables. Use all important covariates or perform variable selections using standard statistics methods.
4. Check for goodness of fit of the models and select the best that fits the data well.
5. Produce residual plots to check for model assumptions including independence/multicollinearity, equal variance, outliers and normality of residuals.
6. In your report, make sure to produce Tables for Descriptive/summary statistics
7. Create a Table(s) of inferential statistics from the final model.
8. Select only a few important graphs (scatterplots/line graphs, boxplots, barchars, etc) and show the relationship between the response and the covariates of interest.
9. Write your full report (in pdf format) and draw conclusions based on your study objective(s).

You are required to make slides from your final report and record your presentation.

Submit both your project report (in pdf format) and the recorded slides/powerpoint presentation for grading.

Qn 1: IPO Dataset

Model Hint/Suggestion: Multiple logistic regression model

Private companies often go public by issuing shares of stock referred to as initial public offerings (IPOs). A study of 482 IPOs was conducted to determine what are the characteristics of companies that attract venture capital funding. The response of interest is whether or not a company was financed with `venture capital funds` (Variable: `funding`). Potential predictors include the `face value of the company (in millions)`, `the number of shares offered (in millions)`, and whether or not the company underwent a `leveraged buyout`. Each line of the data set has an `identification number` and provides information on 4 other variables for a single person. The 5 variables are:

Variable	Variable Name	Description
1	`idnum`: Identification number	$1 - 482$
2	`funding`: Venture capital funding	Presence or absence of venture capital funding: 1 if yes; 0 otherwise
3	`facevalue`: Face value company	Estimated face value of company from prospectus (in Million dollars)
4	`shares`: Number of shares offered	Total number of shares offered (in Millions)
5	`buyout`: Leverage buyout	Presence or absence of leveraged buyout: 1 if yes; 0 otherwise

idnum	funding	facevalue	shares	buyout
1	0	1.2	3	0
2	0	1.45	1.45	1
3	0	1.5	0.3	0
4	0	1.53	0.51	0
...	...	...	...	...
479	0	143.24	11.02	1
480	0	159.5	7.25	0
481	0	165	11	0
482	0	234.6	9.2	0

Qn 2: Prostate Cancer Dataset

Model Hint/Suggestion: Multiple linear regression model

A university medical center urology group was interested in the association between `prostate-specific antigen (PSA level)` is the response variable and a number of prognostic clinical measurements in men with advanced prostate cancer. Data were collected on 97 men who were about to undergo radical prostectomies. Each line of the data set has an `identification number` and provides information on 8 other variables for each person. The 9 variables are:

Table 2: *Adapted in part from: Hastie, T. J.; R. J. Tibshirani; and J. Friedman. The Elements of Statistical Learning: Data Mining. Inference. and Prediction. New York: Springer-Verlag, 2001.*

Variable	Variable Name	Description
1	`idnum`: Identification number	$1 - 972$
2	`psa`: PSA level	Serum prostate-specific antigen level (mg/ml)
3	`cancerv`: Cancer volume	Estimate of prostate cancer volume (cc)
4	`weight`: Weight	Prostate weight (gm)
5	`age`: Age	Age of patient $(years)$
6	`hyperplasia`: Benign prostatic hyperplasia	Amount of benign prostatic hyperplasia (cm^2)
7	`seminal`: Seminal vesicle invasion	Presence Or absence of seminal vesicle invasion: 1 if yes; 0 otherwise
8	`capsular`: Capsular penetration	Degree of capsular penetration (cm)
9	`score`: Gleason score	Pathologically determined grade of disease using total score of two patterns (summed scores were either 6, 7, or 8 with higher scores indicating worse prognosis)

idnum	psa	cancerv	weight	age	hyperplasia	seminal	capsular	score
1	0.65	0.56	15.96	50	0	0	0	6
2	0.85	0.37	27.66	58	0	0	0	7
3	0.85	0.6	14.73	74	0	0	0	7
...	...	...	...	...	...	...	...	...
95	170.72	18.36	29.96	52	0	1	11.7	8
96	239.85	17.81	43.38	68	4.76	1	4.76	8
97	265.07	32.14	52.98	68	1.55	1	18.17	8

Qn 3: Website Developer Dataset

Model Hint/Suggestion: Start with a Poisson regression model, check for over-dispersion, if present, consider fitting a Negative Binomial regression model

Recall that for Poisson regression, one of the assumptions for a valid model is that the mean and variance of the count variable are equal. The negative binomial distribution is a more generalized form of distribution used for 'count' response data, allowing for greater dispersion or variance of counts. In practice, it is quite common for the variance of the outcome to be larger than the mean. This is called overdispersion. If a count variable is overdispersed, Poisson regression underestimates the standard errors of the predictor variables. When overdispersion is evident, one solution is to specify that the errors have a negative binomial distribution.

Management of a company that develops websites was interested in determining which variables have the greatest impact on the `number of websites developed and delivered to customers per quarter` (Response variable: Websites delivered). Data were collected on website production output for 13 three-person website development teams, from January 2001 through August 2002. Each line of the data set has an `identification number` and provides information on 6 other variables for thirteen teams over time. The 8 variables are:

Variable	Variable Name	Description
1	`idnum`: Identification number	$1 - 73$
2	`delivered`: Websites delivered	Number of websites completed and delivered to customers during the quarter
3	`backlog`: Backlog of orders	Number of website orders in backlog at the close of the quarter
4	`teamnum`: Team number	$1 - 13$
5	`experience`: Team experience	Number of months team has been together
6	`change`: Process change	A change in the website development process occurred during the second quarter of 2002: 1 if quarter 2 or 3, 2002; 0 otherwise
7	`year`: Year	2001 or 2002
8	`quarter`: Quarter	$1, 2, 3$, or 4

idnum	delivered	backlog	teamnum	experience	change	year	quarter
1	1	12	1	3	0	2001	1
2	2	18	1	6	0	2001	2
3	7	26	1	9	0	2001	3
4	2	28	1	12	0	2001	4
...	...	...	...	...	...	...	...
70	7	28	13	11	0	2001	4
71	7	36	13	14	0	2002	1
72	19	37	13	17	1	2002	2
73	12	26	13	20	1	2002	3

Qn 4: Market Share Dataset

Model Hint/Suggestion: Multiple linear regression model

Company executives from a large packaged foods manufacturer wished to determine which factors influence the `market share` of one of its products (market share is the response variable). Data were collected from a `national database (Nielsen)` for 36 consecutive months. Each line of the data set has an `identification number` and provides information on 6 other variables for each month. The data presented here are for September, 1999, through August, 2002. The variables are:

Variable	Variable Name	Description
1	`idnum`: Identification number	$1 - 36$
2	`marketshare`: Market share	Average monthly market share for product (percent)
3	`price`: Price	Average monthly price of product (dollars)
4	`gnrpoints`: Gross Nielson rating points	An index of the amount of advertising exposure that the product received
5	`discount`: Discount price	Presence or absence of discount price during period: 1 if discount, 0 otherwise
6	`promotion`: Package promotion	Presence or absence of package promotion during period: 1 if promotion present, 0 otherwise
7	`month`: Month	Month (Jan-Dec)
8	`year`: Year	Year (1999 - 2002)

idnum	marketshare	price	gnrpoints	discount	promotion	month	year
1	3.15	2.2	498	1	1	Sep	1999
2	2.52	2.19	510	0	0	Oct	1999
3	2.64	2.29	422	1	1	Nov	1999
4	2.55	2.42	858	0	1	Dec	1999
...	...	...	...	...	...	NA	...
33	2.88	2.42	145	1	1	May	2002
34	2.8	2.52	270	1	0	Jun	2002
35	2.48	2.5	322	0	1	Jul	2002
36	2.85	2.78	317	1	1	Aug	2002

Qn 5: Disease Outbreak Dataset

Model Hint/Suggestion: Multiple logistic regression model

source = Book Website

Adapted in part from H.G. Dantes, J.S. Koopman, C.L. Addy, et. al., "Dengue Epidemics on the Pacific Coast of Mexico." *International Journal of Epidemiology* 17 (1988), *pp.* 178 – 86

The data set below provides information from a study based on 196 persons selected in a probability sample within two `sectors` in a city. Assume that the response variable (main outcome of interest) is `disease: (Disease status)` which is coded 1 if the person has a disease or 0 if they do not have a disease. Each line of the dat set has an `identification number (id)` and provides information on 5 other variables (exposure/independent variables) for each person. The 6 variables are:

Variable	Variable Name	Description
1	`id`: Identification number	1 – 196
2	`ageyrs`: Age	Age of person (in years)
3	`ses`: Socio-economic status	1 = upper, 2 = middle, 3 = lower
4	`sector`: Sector	Sector within city, where: 1 = sector 1, 2 = sector 2
5	`disease`: Disease status	1 = with disease, 0 = without disease
6	`savings`: Savings account status	1 = has savings account, 0 = does not have savings account

id	ageyrs	ses	sector	disease	savings
1	33	1	1	0	1
2	35	1	1	0	1
3	6	1	1	0	0
4	60	1	1	0	1
...	...	...	...	...	...
193	10	3	1	0	1
194	31	3	1	0	0
195	85	3	1	0	1
196	24	2	1	0	0

Qn 6: Mosquito larva infestation Dataset

Model Hint/Suggestion: Multiple Poisson regression and Negative Binomial regression Models

Recall that for Poisson regression, one of the assumptions for a valid model is that the mean and variance of the count variable are equal. The negative binomial distribution is a more generalized form of distribution used for 'count' response data, allowing for greater dispersion or variance of counts. In practice, it is quite common for the variance of the outcome to be larger than the mean. This is called overdispersion. If a count variable is overdispersed, Poisson regression underestimates the standard errors of the predictor variables. When overdispersion is evident, one solution is to specify that the errors have a negative binomial distribution.

Use the data set `DHF99` from the `R` package `epiDisplay`. Type `library(epiDisplay)` then `?DHF99` to see more details about the dataset.

The main outcome of interest (response variable) is `counts of water containers` infested with mosquito larvae in a field survey. This is variable `containers` in the data.

```
library(epiDisplay)
data("DHF99")
# create a new dataset to manipulate
malaria <- DHF99
summ(malaria)
```

```
##
## No. of observations = 300
##
##   Var. name  obs. mean   median  s.d.   min.   max.
## 1 houseid    300  174.27 154.5   112.44 1      385
## 2 village    300  48.56  51      32.25  1      105
## 3 education  300  2.09   1       1.455  1      5
## 4 containers 299  0.35   0       1.01   0      11
## 5 viltype    300  1.56   1       0.754  1      3
```

```
codebook(malaria)
```

```
##
##
##
## houseid   :   no
##  obs. mean    median  s.d.    min.   max.
##  300  174.273 154.5   112.439 1      385
##
##  ==================
## village   :   Village
##  obs. mean   median  s.d.   min.   max.
```

```
##  300  48.56  51      32.253 1      105
##
##  ==================
## education     :   Educational level
##             Frequency Percent
## Primary           168   56.00
## Secondary          36   12.00
## High school        34   11.33
## Bachelor           25    8.33
## Other              37   12.33
##
##  ==================
## containers    :   # infested vessels
##  obs. mean   median  s.d.   min.   max.
##  299  0.351  0       1.014  0      11
##
##  ==================
## viltype   :   Village type
##       Frequency Percent
## rural       180      60
## urban        72      24
## slum         48      16
##
##  ==================
```

	houseid	village	education	containers	viltype
1	1	22	Other	3	rural
2	2	22	Primary	1	rural
3	3	22	Primary	0	rural
4	4	22	Primary	0	rural
...	...	...	NA	...	NA
297	382	39	Primary	0	rural
298	383	39	Primary	0	rural
299	384	39	Primary	0	rural
300	385	39	Primary	0	rural