

Applied Regression Analysis

.....
Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics, University of
Washington

Session 2

Scientific Setting

.....

Scientific Questions

.....

- Statistics is about science
 - (Science in the broadest sense of the word)

- Science is about proving things to people
 - (The validity of any proof rests solely on the willingness of the audience to believe it)

Scientific Questions

.....

- Inevitably, scientific questions are concerned with investigating cause and effect
 - E.g., in biomedical settings:
 - What are the causes of disease?
 - What are the effects of interventions?

4

Scientific Questions

.....

- Sometimes conditions of scientific studies make answering such questions difficult even when study results are deterministic (no variation in response)
 - Difficulties in isolating specific causes
 - E.g., isolating REM sleep from total sleep
 - E.g., interactions between genetics and environment
 - Difficulties in measuring potential effects
 - E.g., measuring time to survival
 - length of study
 - competing risks

5

Limitations of Statistics

.....

If the scientific question cannot be answered when outcomes are entirely deterministic, there is
NO CHANCE
that statistics can be of any help.

6

Scientific Questions

- Furthermore, there is inevitably variation in response across repetitions of an experiment
 - Variation can be due to
 - Unmeasured (hidden) variables
 - E.g., mix of etiologies, duration of disease, comorbid conditions, genetics when studying new cancer therapies
 - Inherent randomness
 - (as dictated by quantum theory)

7

Scientific Questions

- Scientific questions thus have to be phrased in a manner that acknowledges such variation in response
 - Deterministic:
 - Does meditation decrease blood pressure?
 - Probabilistic:
 - Does meditation tend to decrease blood pressure?

8

Scientific Questions

- Of course, the probabilistic approach only makes sense if we can find a suitable definition for the phrase “tends to”
 - Many possibilities exist for detecting a decrease:
 - A lower average value (arithmetic mean)
 - A lower geometric mean
 - A lower median: $Mdn (Trt) - Mdn (Ctrl) < 0.0$
 - Median (Treated – Control) < 0.0
 - A lower proportion exceeding some threshold
 - A lower odds of exceeding some threshold
 - $Pr (Treated > Control) < 0.5$
 - Time average of hazard ratio < 1.0

9

Scientific Questions

.....

- The choice of the definition of “tends to” should be dictated first by scientific considerations
 - E.g., Is the arithmetic mean’s sensitivity to outliers desirable or undesirable?
 - Does making one person immortal make up for killing others prematurely?
 - E.g., Is the scientific importance of a difference in distribution best measured by the proportion exceeding some threshold?
 - Is a decrease in blood glucose levels in diabetes only important if normal levels are attained?

10

Limitations of Statistics

.....

If the scientific researcher cannot decide which parameters would be appropriate when measurements are available on the entire population, there is
NO CHANCE
that statistics can be of any help. ¹¹

Role of Statistics

.....

12

Statistical Tasks

- Statistics plays a role at multiple stages of the conduct of a scientific study
 - Refine the scientific question
 - Study design
 - Descriptive statistics
 - Inferential statistics
 - Computing estimates of population parameters
 - Quantifying strength of evidence in data

13

Classification of Statistical Questions

1. Prediction of individual observations
2. Identifying clusters of observations
3. Identifying clusters of variables
4. Quantifying the distribution of some variable
5. Comparing the distributions of some variable across groups
 - Identifying groups having different distributions of response
 - Associations between response and grouping variables
 - Effect Modification: Differences of associations

14

Questions Answered with Regression

1. Prediction of individual observations

4. Quantifying the distribution of some variable
5. Comparing the distributions of some variable across groups
 - Identifying groups having different distributions of response
 - Associations between response and grouping variables
 - Effect Modification: Differences of associations

15

Questions Answered with Regression.....

1. Prediction of individual observations
4. Quantifying the distribution of some variable
5. Comparing the distributions of some variable across groups
 - Identifying groups having different distributions of response
 - Associations between response and grouping variables
 - Effect Modification: Differences of associations

16

Example: Normal Range for SEP.....

- Diagnosis of demyelinating disease
 - Some diseases of the peripheral nervous system are associated with slower transmission of nerve signals
 - Somatosensory evoked potentials (SEP) are one means of detecting slower transmission
 - Stimulate a nerve in the foot
 - Measure time until an impulse is detected in the brain (EEG)

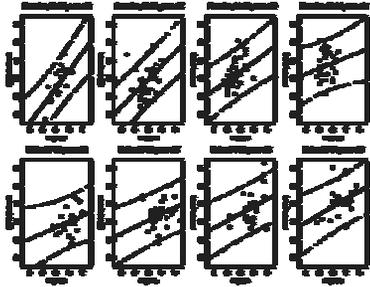
17

Example: Normal Range for SEP.....

- Scientific question
 - What are the normal ranges for SEP?
 - (Do they depend on height, age, sex?)
- Statistical analysis
 - SEP measured in a sample of healthy volunteers
 - Regression modeling mean SEP linear in height
 - (Line specific to each age, sex)
 - 95% prediction intervals to define “normal” range
 - Approximately: Mean \pm 2 SD

18

Example: Normal Range for SEP



19

Example: Normal Range for SEP

- Conclusion
 - “Normal” ranges can be provided for each combination of height, age, and sex
 - How should these intervals be communicated?
 - Is the assumption of the normal distribution valid?

20

Questions Answered with Regression

- Prediction of individual observations

- Quantifying the distribution of some variable
- Comparing the distributions of some variable across groups
 - Identifying groups having different distributions of response
 - Associations between response and grouping variables
 - Effect Modification: Differences of associations

21

Example: Prognosis in PBC

- Median residual life expectancy in Primary Biliary Cirrhosis (PBC)
 - Primary biliary cirrhosis is a severe, idiopathic disease of the liver
 - Progression of disease often leads to liver transplantation

22

Example: Prognosis in PBC

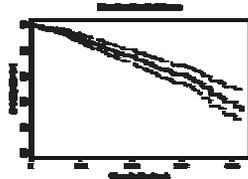
- Scientific question
 - Which patients should be transplanted within a year?
 - (We want to list for transplant with adequate time)
- Statistical analysis
 - Follow a cohort of 418 patients with PBC
 - Proportional hazards regression model of time to death
 - Risk of death modeled by bilirubin, edema, etc.
 - Model used to predict median time to death

23

Example: Prognosis in PBC

- Proportional hazards model

$$S(t | \bar{Z}) = S_0(t)^{e^{\bar{Z}\beta}}$$



	Coef	SE(Coef)	P value
log(bili)	0.9434	0.08093	0.0e+000
edema	0.8915	0.19520	4.9e-006
log(prottime)	2.5255	0.78549	1.3e-003
age	0.0397	0.00762	2.0e-007

24

Example: Prognosis in PBC

- Conclusion
 - “Mayo R-score” is a standard index of risk in liver studies
 - Baseline survival probabilities tabled in publication
 - Conceptually, inference about the median survival time could also be obtained from this model

25

Questions Answered with Regression.....

- Prediction of individual observations

- Quantifying the distribution of some variable
- Comparing the distributions of some variable across groups
 - Identifying groups having different distributions of response
 - Associations between response and grouping variables
 - Effect Modification: Differences of associations

26

Example: Predictors of Cerebral Atrophy

- MRI measurements of cerebral atrophy
 - Patterns of cerebral changes seen on brain MRI in older people
 - Widening of sulci, ventricles
 - Clinical importance of such changes is unclear

27

Example: Predictors of Cerebral Atrophy

- Scientific question
 - What clinical and subclinical characteristics are associated with cerebral atrophy?
- Statistical analysis
 - Measure a cohort of ~3,000 elderly people
 - Stepwise linear regression to identify list of most promising characteristics for further study
 - "Significant" predictors from a preliminary list of 23 variables measuring demographics, concurrent disease

28

Example: Predictors of Cerebral Atrophy

- Linear regression model of mean atrophy score
 - Scaled 0 (least atrophy) to 100 (best atrophy)

$$E(\text{Atrophy}) = \beta_0 + X_1\beta_1 + \dots + X_p\beta_p$$

Resid Std Err = 11.9214, Mult R-Square = 0.1525
N= 735, F-stat= 43.838 on 3, 731 df, P-val< .0001

	coef	std.err	t.stat	p.value
Intercept	-16.8057	6.0486	-2.7784	0.0056
Age	0.6620	0.0810	8.1756	0.0000
Male	5.7222	0.8829	6.4810	0.0000
Stroke	4.2594	1.2963	3.2857	0.0011

29

Example: Predictors of Cerebral Atrophy

- Conclusions
 - Of the 23 potential risk factors, only age, sex, and history of stroke were identified
 - Approximately equivalent difference in mean atrophy scores for
 - 10 year difference in age
 - History of stroke versus no history of stroke
 - Male versus females
 - We cannot trust the P values, because of the multiple testing used to select the model
 - Further studies are necessary to confirm these estimates

30

Multiple Comparison Problem

.....
“When you go looking for something specific, your chances of finding it are very bad, because of all the things in the world, you’re only looking for one of them.

“When you go looking for anything at all, your chances of finding it are very good, because of all the things in the world, you’re sure to find some of them.”

- Darryl Zero in “The Zero Effect”

Multiple Comparison Problem

.....
- “When you go looking for something specific, your chances of finding [a spurious association by chance] are very bad, because of all the things in the world, you’re only looking for one of them.

- “When you go looking for anything at all, your chances of finding [a spurious association by chance] are very good, because of all the things in the world, you’re sure to find some of them.”

Questions Answered with Regression

.....
- Prediction of individual observations

- Quantifying the distribution of some variable
- Comparing the distributions of some variable across groups
 - Identifying groups having different distributions of response
 - Associations between response and grouping variables
 - Effect Modification: Differences of associations

Example: Smoking Effect on FEV

- Association between smoking and lung function in children
 - Long term smoking is associated with lower lung function
 - Are similar effects observed in short term smoking in children?

34

Example: Smoking Effect on FEV

- Scientific question
 - Does smoking lead to lower lung function in kids?
- Statistical analysis
 - Measure smoking behavior, lung function in 654 healthy children
 - Compare geometric mean of FEV across groups
 - Smokers versus nonsmokers
 - Smokers versus nonsmokers of same age

35

Example: Smoking Effect on FEV

- Linear regression model of log FEV (compares ratio of geometric means across groups)
 - Unadjusted Analysis

Resid Std Err= 0.2477, Multiple R-Square= 0.0212
N= 439, F-stat= 9.4501 on 1, 437 df, P-val= 0.0022

	coef	std.err	Geom Mn	t.stat	p.value
Intercept	1.0582	0.0128	2.8811	82.632	0.0000
Smoker	0.1023	0.0333	1.1077	3.074	0.0022

36

Example: Smoking Effect on FEV

- Linear regression model of log FEV (compares ratio of geometric means across groups)
 - Age and height adjusted Analysis

Resid Std Err= 0.1441, Multiple R-Square= 0.6703
N = 439, F-stat = 294.73 on 3, 435 df, P-val< .0001

	coef	std.err	Geom Mn	t.stat	p.value
Intercept	-11.0946	0.5201	0.0000	-21.3306	0.0000
Smoker	-0.0536	0.0209	0.9478	-2.5584	0.0109
Age	0.0215	0.0038	1.0218	5.6379	0.0000
LogHT	2.8697	0.1301	17.6310	22.0645	0.0000

37

Example: Smoking Effect on FEV

- Conclusions
 - We find differences in distribution of lung function between children who say they smoke and those who say they do not
 - Children who smoke tend to have FEV approximately 11% higher than children who do not smoke
 - Children who smoke tend to have FEV approximately 5% lower than children of the same age and height who do not smoke
 - These differences are atypical of chance observations made in the absence of true associations
 - $P = 0.0022$ and $P = 0.0109$, respectively

38

Questions Answered with Regression

- Prediction of individual observations
- Quantifying the distribution of some variable
- Comparing the distributions of some variable across groups
 - Identifying groups having different distributions of response
 - Associations between response and grouping variables
 - Effect Modification: Differences of associations

39

Example: Age-Smoking Interaction

- Interaction between age and smoking on lung function in children
 - Is the effect of smoking on FEV the same in all age groups?

40

Example: Age-Smoking Interaction

- Scientific question
 - Is the effect of smoking on FEV the same in all age groups?
- Statistical analysis
 - Include a variable modeling the age-smoking interaction in the linear regression of log FEV

41

Example: Age-Smoking Interaction

- Linear regression model of log FEV (compares ratio of geometric means across groups)
 - Analysis with interaction

Resid Std Error = 0.144, Multiple R-Square = 0.6712
N = 439, F-stat = 221.48 on 4, 434 df, p-value = 0

	coef	std.err	Geom Mn	t.stat	p.value
Intercept	-10.9963	0.5274	0.0000	-20.8488	0.0000
Smoker	0.0731	0.1157	1.0758	0.6318	0.5279
SmokAge	-0.0097	0.0087	0.9904	-1.1133	0.2662
Age	0.0238	0.0043	1.0240	5.5130	0.0000
LogHT	2.8400	0.1327	17.1162	21.3989	0.0000

42

Example: Age-Smoking Interaction

- Conclusions
 - We do not find a difference in the smoking effect on FEV across different age groups beyond that which could be explained by coincidental observations
 - $P = 0.2662$
 - (I am usually hesitant to interpret interaction parameter estimates too closely: We probably did not model the interaction very well)

43

Analysis of Data

.....

44

Analysis of Data

- Statistical analysis of the data
 - Descriptive statistics which summarize the sample
 - Inferential statistics which allow generalization to the population from which the sample was drawn
 - Provide estimates
 - Quantify the precision of the estimates
 - Aid in decisions

45

Purpose of Descriptive Statistics

.....

- Summarize a large quantity of data in a few statistics that capture the essence of the entire data set.
 - The best descriptive statistic for a particular problem depends upon the data
 - Choose the measures that summarize the data best (with the fewest statistics) for the purpose at hand

46

Purpose of Descriptive Statistics

.....

- Detect errors in the data
 - Missing data
- Describe the materials and methods used
 - Sampling scheme
- Assess conditions that might alter analysis methods
 - Modeling interactions (effect modification)
 - Adjusting for confounding
 - Adjusting for variables that provide precision
- Provide estimates that address scientific questions
- Generate new hypotheses

47

Purpose of Inferential Statistics

.....

- It is rare that anyone would analyze data and not be interested in drawing inference from it.
 - I would therefore argue that quantifying our uncertainty is always appropriate.
 - However:
 - all our methods are dependent upon assumptions
 - some methods are more dependent upon assumptions than others
 - no statistics can correct for inappropriately gathered data

48

Purpose of Inferential Statistics

- Types of inference
 - Best estimate of what is true in the population
 - Quantify the strength of evidence in the sample
 - Range of “reasonable” estimates
 - Probability reflecting “confidence” in decisions
 - Binary decision based on one of the above

49

Scientific versus Statistical Measures

- A common pitfall of data analysis is to believe that statistical measures of precision capture all the relevant information
 - I find it extremely important to distinguish between
 - Scientific estimates of treatment effect which allow us to judge the clinical relevance of an association, and
 - Statistical estimates of study precision which allow us to quantify our confidence in the study results.

50

An Aside: Characterizing Associations

- Hypothetical study to detect an association between Event B and Exposure F
 - Unexposed: 0 of 5 have Event B
 - Estimated incidence rate: 0.000
 - 95% CI for incidence rate: 0.000 – 0.522
 - Exposed: 3 of 5 have Event B
 - Estimated incidence rate: 0.600
 - 95% CI for incidence rate: 0.147 – 0.947
 - Fisher’s Exact two-sided P: 0.167
 - How would you characterize the presence of an association between these two variables?

51

An Aside: Characterizing Associations.....

- **WRONG:** An overstated, purely statistical report
 - “As the P value is greater than 0.05, we conclude that there is no association between exposure F and event B.”
 - (We should not conclude that there is no association, because we lacked precision to rule out differences that might be of interest.)

52

An Aside: Characterizing Associations.....

- **TECHNICALLY CORRECT, BUT OF LITTLE USE:** A correctly stated, purely statistical report
 - “As the P value is greater than 0.05, we conclude that there is not sufficient evidence to rule out the possibility that there is no association between exposure F and event B.”
 - (This is stated correctly, but it gives us no idea whether we had ruled out differences that we cared about or we had merely done an abysmal study.)

53

An Aside: Characterizing Associations.....

- **CORRECT AND SCIENTIFICALLY USEFUL:** A correctly stated report of scientific estimates and quantification of statistical evidence
 - “We observed incidence rates of 60% in the exposed (95% CI: 15% - 95%) and 0% in the unexposed (95% CI: 0% - 52%). Unfortunately, the precision of the study was not adequate to demonstrate that such a large difference in incidence rates would be unlikely in the absence of a true association (P = 0.17).”

54

An Aside: Characterizing Associations.....

- For what it is worth...
 - These data are not atypical of what we might expect to see if F= female and B= giving birth
 - Had I told you that at the start, very few would have interpreted the P values so concretely
 - We already know an awful lot about the association between sex and pregnancy
 - However, in real science, we are studying questions we do not know the answer to

55

Regression.....

- The primary statistical methods that we will use are those referred to as regression
 - Regression methods differ primarily according to the parameter compared across groups
 - Means → Linear regression
 - Geom Means → Linear regression on logs
 - Odds → Logistic regression
 - Rates → Poisson regression
 - Hazards → Proportional Hazards regression
 - Quantiles → Parametric survival regression

56

How It Is All Possible.....

- Most of our inferential methods have their root in the central limit theorem
 - Sample means are asymptotically normally distributed across (conceptual) replications of studies
 - Properties of the normal distribution allow us to
 - average statistics across groups (adjust)
 - compare statistics between groups (detect associations)

57

Basic Approach

- Because we tend to use asymptotic normal statistics
 - Population parameters are estimated by corresponding sample descriptive statistic
 - Confidence intervals usually of the form
 - estimate \pm std err \times crit value
 - Tests performed by standardization and comparison to a critical value
 - $(\text{estimate} - \text{null value}) / (\text{std err})$
 - (Critical value approximately 2 for a 95% CI or a two-sided 0.05 test)

58
