Biost 518 / Biost 515
Applied Biostatistics II / Biostatistics II

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

Lecture 12:

Correlated Response; Weighted Regression

March 6, 2015

1

Lecture Outline

- Dependent data within clusters
- Weighted regressions

2

Dependent Data
Within Clusters

3

Dependent Data

- There are times when data can not be presumed to be totally independent
  - Sampling within families
  - Sampling within schools, hospitals
  - Repeated measurements on individuals taken at a single time
  - Longitudinal data: repeated measurements taken on individuals over time

4

Motivation for Longitudinal Data

- Three settings in which longitudinal studies are performed
- Convenience of existing study population
- Efficiency
  - Repeated measurements to decrease variability
  - Using subjects as own comparison
- Scientific questions about effects that occur
  - over time, or
  - within subjects

5

Convenience

- Questions are truly cross-sectional
- Multiple measurements made on each individual is easier than gathering new subjects
  - Natural variation within individuals provides additional information
- E.g., Serum osmolality from Na, Glc, BUN
    - Interest is relationships between concurrent measurements

6

Efficiency

- Questions could be answered with cross-sectional study
- Primary comparison within subjects may have less variability
  - Allow detection of smaller effects
  - E.g., Adjusting for baseline measurements
  - E.g., Cross-over study of a new treatment

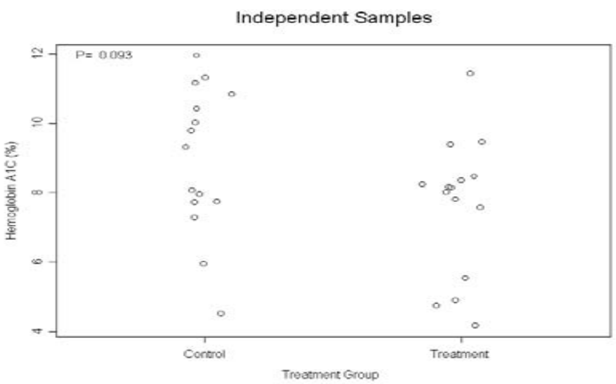
7

Example

- Percent glycosylated hemoglobin is used to monitor long term control in diabetes
  - Hemoglobin A1c
- Consider studies of two insulin delivery strategies
  - Independent groups
  - Cross-over design

8

Graph: Independent Samples



9

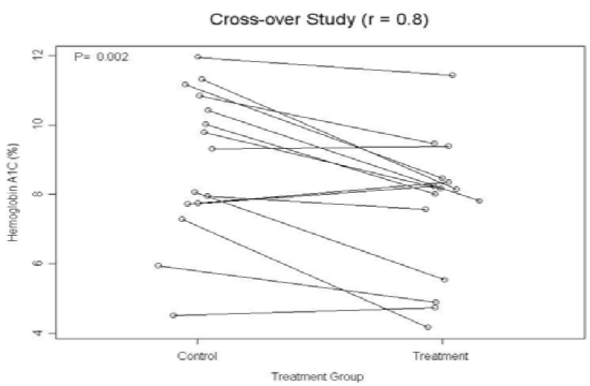
Inference: Independent Groups

- Large between-subject variability hampers our ability to detect differences
  - Between group SE is square root of sum of squared within group SEs
  - Within group SEs are proportional to within group standard deviation divided by the square root of n

$$se\left(\overline{X} - \overline{Y}\right) = \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$$

10

Graph: Cross-over Study



11

Inference: Cross-over Study

- High correlation between measurements taken on the same individual increases precision
  - The "random effect" of patient ID can be thought of as a precision variable

$$se\left(\overline{X} - \overline{Y}\right) = se\left(\overline{D}\right) = \sqrt{\frac{\sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y}{n}}$$

12

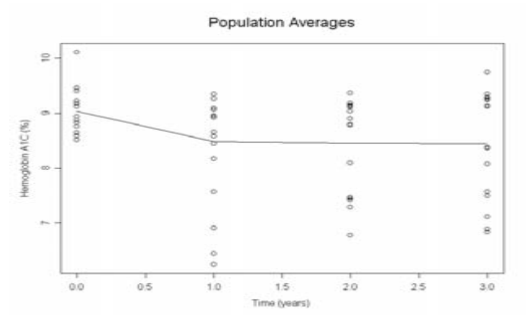
Longitudinal Questions

- Scientific questions about effects that occur over time
  - Studies to detect population time trends in response
    - E.g., rate (slope) of progression of retinopathy in population of diabetics over time
    - E.g., time to development of albuminuria

13

Example: "Marginal Effects"

- Time trends in group mean HbA1C
  - Note trends in mean and variability



14

Within Subject Effects

- Trends in specific individuals might not look like trends in population means
- Response over time may be restricted to subgroups of subjects
- Response over time may be transient

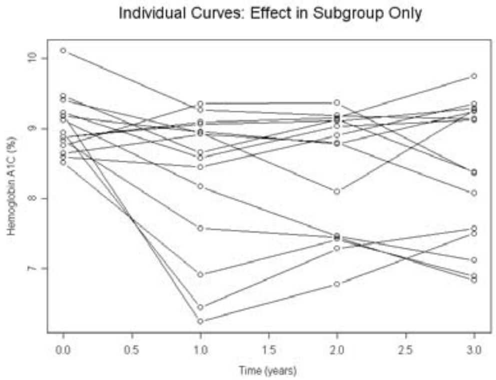
15

Longitudinal Scientific Questions

- Scientific questions about effects that occur within subjects
  - Studies to detect time trends or covariate effects in individual response
    - E.g., distribution of rates (slopes) of progression of retinopathy in population over time
    - E.g., effect of varying risk factors within individuals

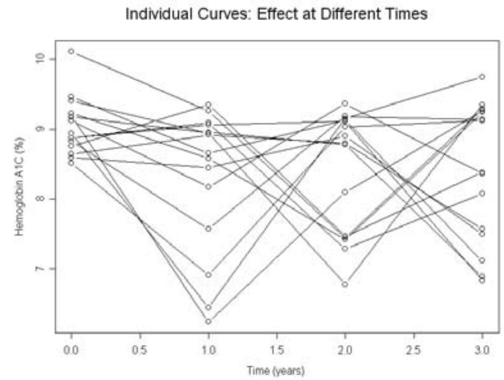
16

Effect in Subgroup



17

Transient Effects



18

Choice of Measures of Outcome

- In order of importance
  - Scientific relevance
    - Including state of current knowledge
  - Plausibility of difference across groups
  - Statistical precision for analysis

19

Longitudinal Outcome Measures

- In longitudinal studies, each individual may have multiple measurements over time
- Definition of individual response thus can be based on multiple measurements
  - Response at a fixed time
  - Responses at multiple fixed times
  - Average response over time (area under curve)
  - Rate of change in response (slope)
  - Time to attaining some level of response

20

Measures of Outcome

- "Marginal" or population effects
  - Difference or ratio of group means, geometric means, medians, proportion or odds above threshold, hazards
  - Pr (Y > X)
- "Within subject" effects
  - Mean, median difference
  - Mean, geometric mean, median ratio
  - Within subject odds ratio
  - Pr (Y > X)

21

Choice of Longitudinal Outcome

- Should reflect scientific relevance, plausibility of effect, precision
- Final level of response may be more important than earlier effects
  - (But in the long run, we are all dead)
- Summarizing response at multiple time points reflects population rather than individuals
- Average response over time sensitive to transient effects
- Differences in time to event may be clinically meaningless

22

Statistical Issues

- Repeated measurements on subjects require special analysis techniques
- May have erroneous conclusion if fail to account for correlated observations
  - Point estimates may be biased for population parameters
    - Too much emphasis placed on some subjects
  - Confidence intervals will not be accurate representation of our true confidence
  - P values will be wrong

23

Statistical Approaches

- Three basic approaches to analyzing correlated data
- Reduce measurements on each cluster to a single observation; analyze across clusters
- Estimate correlation within clusters and adjust standard errors for population based models
  - GEE, marginal models
  - "Robust" variance estimates
- Adjust estimates for "random effects"
  - "Mixed effects models": both fixed and random

24

Easiest Approach

- Reduce data for each individual to a single measurement
  - E.g., response at end of study, average response, rate of change
  - Analyses can then be based on standard methods for independent data
  - But:
    - Does not allow time-varying covariates
    - May not be most efficient statistically

25

Example: Beta-carotene Data

- Randomized clinical trial of beta-carotene supplementation on plasma levels of beta-carotene and vitamin E
  - Subjects randomized to 5 dose groups
  - Measurements at baseline, after 3 and 9 months of treatment, and 3 months after stopping treatment
  - Scientific question: How do plasma beta-carotene levels change over time within dose groups?
    - (effect modification between dose and time)

26

Example: Beta-carotene Data

- Reduce data to a single measurement on each subject
- Difference between follow-up and baseline
  - Consider average of differences
  - No change corresponds to a difference of 0
- Ratio between follow-up and baseline
  - Consider average of ratios
  - No change corresponds to a ratio of 1

27

Example: SEP data

- Somatosensory evoked potential measurements on healthy adults
- Measurements of nerve conduction time
  - Four separate peaks for each leg of each subject
- Reduce data to a single measurement
  - Consider only one peak on one leg
    - Which one?
  - Average measurements across peaks, legs
    - But will only generalize to similar averages
  - (Differences between peaks?)

28

Two Matched Samples

- Paired t test for means
- Sign test for median difference
- Wilcoxon signed rank test
- McNemar's test for difference in proportions or odds ratio
  - Equivalent to sign test

29

General Approach

- Regression models which allow correlated data within identified clusters
- Allows time-varying covariates
- Allows greater precision in some settings
  - Recall increased precision by adjusting for baseline as a predictor in RCT

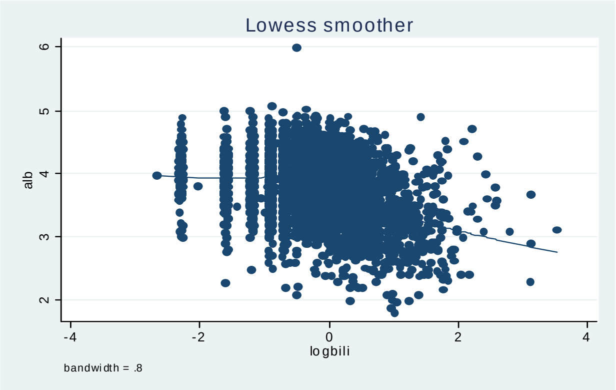
30

Example: Bilirubin & Albumin

- How do albumin levels relate to bilirubin levels in PBC?
- Randomized clinical trial
  - 265 subjects
  - Measurements made every 3 months
    - 9,068 measurements
- Transform to log bilirubin for skewness

31

Lowess Smooth: Alb vs Logbili



32

Regression Results

```
. regress alb logbili, robust
Linear regression          Number of obs =    9068
                           F(  1,  9066) =  464.37
                           Prob > F      =  0.0000
                           R-squared     =  0.0729
                           Root MSE      =  .39451
-------------------------------------------------
        |       Robust
    alb |  Coef    SE       t   P>|t|    [95% C I]
-------------------------------------------------
logbili | -.171 .0080  -21.55 0.000 -.187  -.156
  _cons |  3.82 .0060  635.72 0.000  3.81   3.83
```

33

So far: Inferential Assumptions

- There are three basic assumptions for regression based tests of associations
- Independence
  - Independent observations
- Variance
  - Linear regression: Equal variance or robust standard errors
  - Logistic, Poisson, PH: Correct model or robust standard errors
- Normally distributed estimates
  - Large sample size
    - Linear: If normally distributed data within groups, 2 observations are “large”
    - Binary, count, or survival data: Large number of events

34

Now: Inferential Assumptions

- There are three basic assumptions for regression based tests of associations
- Independence
  - Independent observations ***between clusters***
- Variance
  - Linear regression: Equal variance or robust standard errors
  - Logistic, Poisson, PH: Correct model or robust standard errors
- Normally distributed estimates
  - Large sample size
    - Linear: If normally distributed data within groups, 2 observations are “large”
    - Binary, count, or survival data: Large number of events

35

Adjusting for Correlated Data

- In order to estimate the true standard errors need to account for correlated data
- Adjustment will depend on
  - Whether statistic involves sums or differences of correlated observations
  - Whether data are positively or negatively correlated

36

Effect of Correlated Data

- If statistic adds correlated observations
  - Positively correlated data leads to larger SE than independent data
- If statistic subtracts correlated observations
  - Positively correlated data leads to smaller SE than independent data

$$Var\left(Y_i + Y_j\right) = Var\left(Y_i\right) + Var\left(Y_j\right) + 2\rho\sqrt{Var\left(Y_i\right)Var\left(Y_j\right)}$$
$$Var\left(Y_i - Y_j\right) = Var\left(Y_i\right) + Var\left(Y_j\right) - 2\rho\sqrt{Var\left(Y_i\right)Var\left(Y_j\right)}$$

37

Example: "Repeated Measures"

- Subjects in a group have multiple measurements
- E.g., Average blood pressure by sex
  - Repeated measurements on subjects who are always in the same group
    - Likely tend to be positively correlated
  - Repeated measurements added to get mean
  - Failure to account for correlated data will tend to underestimate true SE
    - "Anti-conservative" inference: P values too low, CI too narrow

38

Example: "Crossover"

- Subjects contribute information to different groups
- E.g., RCT with each subject on placebo and active treatment
  - Positively correlated measurements subtracted when computing difference in group means
  - Failure to account for correlated data will tend to overestimate true SE
    - "Conservative" inference
    - P values too high, CI too wide

39

Example: Combination

- Regression sometimes adds, sometimes subtracts correlated observations
  - Depends on whether predictors for repeated measures are larger or smaller than mean

$$\text{Model} \quad : \quad E\left[Y_i \mid X_i\right] = \beta_0 + \beta_1 \times X_i$$

$$\text{Estimate} \quad : \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i\left(X_i - \overline{X}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

  - Failure to account for correlated data can lead to either conservative or anti-conservative inference

40

Important: Accounting for Correlation

- The major criterion to use when deciding to adjust for possible correlation is according to your beliefs about possible correlation
  - But it is correlation among observations ***after*** adjustment for any covariates that matters
- Although it is tempting to try to evaluate correlation among the observations, we cannot in general rely on estimating the correlation to make such decisions
  - Even very small correlations can make a big difference when there are many correlated observations within a cluster

41

Illustration: Clustered Sampling

- Consider the standard error of the sample mean of potentially correlated observations

Sample M subject in K independent clusters

Data for ith subject in cluster k: $\quad Y_{ki} \sim \left(\mu, \sigma^2\right)$

Correlation within clusters: $\quad corr\left(Y_{ki}, Y_{kj}\right) = \rho$

Correlation between clusters $k \neq l$: $\quad corr\left(Y_{ki}, Y_{lj}\right) = 0$

Distribution of sample mean: $\quad \bar{Y} \sim N\left(\mu, \frac{\sigma^2}{MK}\left(1+(M-1)\rho\right)\right)$

  - If correlation is truly 0, then denominator of standard error is total number of observations
  - But if size of clusters is large, even a small correlation has a big effect on the standard error
    - For example, *M=100* with *ρ= .01* doubles the standard error relative to independent data

42

Illustration: Clustered Sampling

- Inflation of type 1 error when sampling 1000 subjects
  - Cluster size *M* = 1 – 100; Correlation within clusters *ρ* = 0 to 0.5
  - (Note: Power to detect nonzero correlation of .05 with 500 clusters is approximately 8%)

	Correlation ρ						
M	0	0.01	0.02	0.05	0.1	0.2	0.5
1	0.025	0.025	0.025	0.025	0.025	0.025	0.025
2	0.025	0.026	0.027	0.031	0.037	0.051	0.096
5	0.025	0.030	0.035	0.051	0.081	0.138	0.257
10	0.025	0.036	0.048	0.088	0.151	0.242	0.361
20	0.025	0.050	0.078	0.157	0.250	0.342	0.426
50	0.025	0.094	0.161	0.285	0.370	0.428	0.469
100	0.025	0.162	0.255	0.371	0.429	0.462	0.485

43

Adjusting for Correlation

- Huber – White “sandwich” estimator
  - “Robust standard error estimates”
    - Adjust for unequal variances
    - Adjust for correlation within identified clusters
  - Does not change the regression parameter estimates

44

Stata: Adjusting for Correlation

- With any regression command use option
  - "…, `cluster(varname)`"
  - The cluster variable is usually nominal
  - Can also specify "`robust`", but Stata assumes that robust SE should be used whenever clusters are identified

45

Presuming Independence

```
. regress alb logbili, robust
Linear regression             Number of obs =    9068
                              F(  1,  9066) =  464.37
                              Prob > F      =  0.0000
                              R-squared     =  0.0729
                              Root MSE      =  .39451
         |       Robust
     alb |  Coef    SE       t    P>|t|    [95% C I]
logbili | -.171 .0080  -21.55 0.000 -.187  -.156
  _cons |  3.82 .0060  635.72 0.000  3.81   3.83
```

46

Adjusting for Clusters

```
. regress alb logbili, robust cluster(ptid)
Linear regression             Number of obs =    9068
                              F(  1,   264) =   38.29
                              Prob > F      =  0.0000
                              R-squared     =  0.0729
No. clusters (ptid) = 265    Root MSE      =  .39451
        |       Robust
    alb |  Coef    SE       t    P>|t|    [95% C I]
logbili | -.171 .0277   -6.19 0.000 -.226  -.117
  _cons |  3.82 .0235  162.19 0.000  3.77   3.86
```

47

Important Caveat

- This approach does not alter the regression parameter estimates
- By default, each observation is weighted equally
  - Science:
    - Perhaps each subject should be weighted equally
  - Statistics:
    - Perhaps more precise measurements should be weighted more heavily

48

Regression on Binary Predictors

49

Means: Linear Regression

- Classical regression
  - t test which presumes equal variance
- Robust standard errors
  - t test which allows unequal variance (approx)
- Robust standard errors with identified cluster variable
  - paired t test (approx)

50

Example: Classical LR

- In PBC placebo patients, a comparison of mean bilirubin levels between patients with and without edema
  - Standard variance estimates: Compare
    - "`ttest bili, by edema`", and
    - "`regress bili edema`"

51

Standard Variance Estimates

```
. ttest bili if treatmnt==2, by(edema)
Group| Obs     Mean     StErr.    StDev.       [95% CI]
   0 | 137    3.018      .387     4.535     2.252     3.784
   1 |  16    9.25      1.941     7.764     5.113    13.387
diff |       -6.232     1.308              -8.816    -3.647
Ha: diff < 0     Ha: diff ~= 0               Ha: diff > 0
t =  -4.764      t =  -4.764                  t =  -4.764
P<t= 0.0000      P> |t|= 0.0000              P > t = 1.000


. regress bili edema if treatmnt==2
      |           Robust
 bili | Coef.   StErr.      t     P>|t|          [95% CI]
edema | 6.232   1.308      4.76    0.000      3.647     8.816
_cons | 3.018    .423      7.14    0.000      2.182     3.854
```

52

Correspondence

- Estimated intercept: the group 0 sample mean
  - Standard errors differ, because standard regression assumes equal variance between the groups and uses a pooled estimate of variance
  - CI will therefore also be different
- Estimated slope: the difference between group sample means
  - Standard error for slope is the standard error for the difference in means
  - CI, t statistic, and P value are exactly the same

53

Example: LR w/ Robust SE

- In PBC placebo patients, a comparison of mean bilirubin levels between patients with and without edema
- Robust variance estimates: Compare
  - "`ttest bili, by(edema) unequal`", and
  - "`regress bili edema, robust`"

54

Robust Variance Estimates

```
. ttest bili if treatmnt==2, by(edema) unequal
Group| Obs     Mean    StErr.   StDev.       [95% CI]
   0 | 137     3.018    .387    4.535     2.252     3.784
   1 |  16     9.25    1.941    7.764     5.113    13.387
diff |        -6.232   1.979             -10.423    -2.040
Ha: diff < 0     Ha: diff ~= 0              Ha: diff > 0
t =  -3.1484     t =  -3.1484              t =  -3.1484
P < t = 0.0031   P > |t| = 0.0061          P > t = 0.9969


. regress bili edema if treatmnt==2, robust
      |           Robust
 bili | Coef.   StErr.     t    P>|t|         [95% CI]
edema | 6.232   1.931     3.23   0.002     2.416     10.048
_cons | 3.018    .389     7.77   0.000     2.250      3.786
```

55

Correspondence

- Estimated intercept: the group 0 sample mean
  - Standard errors agree more here
  - CI differs a bit more, because of difference in degrees of freedom: t test used 16, regression used 151
- Estimated slope: the difference between group sample means
  - Standard error for slope is approximately the standard error for the difference in means
  - CI, t statistic, and P value are about the same but again are influenced by degrees of freedom

56

Effect of Heteroscedasticity

- Note the problem with using standard linear regression in the presence of unequal variances
- Distribution of predictor (edema) was skewed toward the group with the lower variance
  - Sample size largest in group with smaller variance
- We thus expect standard error estimates from classical regression to be too low
  - Anti-conservative inference

57

Example: Classical LR (Wrong)

- In beta carotene datasets, a comparison of mean vitamin E levels at baseline and after 3 months of treatment (ignoring dose for the purpose of this illustration)
- Standard variance estimates: Compare
  - "`ttest vite0=vite1`", and
  - "`regress vite time`" on reshaped data

58

Classical LR on All Data

```
ttest vite0=vite1
Varbl | Obs    Mean    StErr.   StDev.          [95% CI]
vite0 |  45    8.025    .191    1.280    7.640    8.409
vite1 |  45    8.859    .135     .907    8.586    9.131
diff  |  45    -.834    .133     .894   -1.103    -.566
Ha: mn(diff) < 0   Ha: mn(diff) ~= 0   Ha: mn(diff) > 0
t =  -6.2598       t =  -6.2598        t =  -6.2598
P < t= 0.0000      P > |t|= 0.0000     P > t= 1.0000

regress vite time if time==0 | time==1
                       Number of observations =    91
 vite | Coef.   StErr.       t    P>|t|      [95% CI]
 time |  .799   .234     3.41   0.001     .334  1.264
_cons | 8.060   .165    48.98   0.000    7.733  8.387
```

59

Correspondence

- In this example, no correspondence between the two methods of analysis
- Analyses were performed on different datasets
  - Case 40 did not have measurements at 3 months
  - The paired t test did not use any part of that subject's data
  - The regression used the data at baseline
  - (We can restrict the regression to use the same data)

60

Ex: Classical LR on Same Data

```
. ttest vite0=vite1
Varbl | Obs    Mean    StErr.   StDev.           [95% CI]
vite0 |  45    8.025    .191    1.280    7.640    8.409
vite1 |  45    8.859    .135     .907    8.586    9.131
diff  |  45    -.834    .133     .894   -1.103    -.566
Ha: mn(diff) < 0   Ha: mn(diff) ~= 0   Ha: mn(diff) > 0
t =  -6.2598       t =  -6.2598        t =  -6.2598
P < t= 0.0000      P > |t|= 0.0000     P > t= 1.0000

. regress vite time if (time==0 | time==1) & ptid!=40
                       Number of observations =    90
 vite | Coef.   StErr.       t     P>|t|     [95% CI]
 time |  .834   .234      3.57    0.001     .370  1.299
_cons | 8.025   .165     48.54    0.000    7.696  8.353
```

61

Correspondence

- Estimated intercept: the group 0 sample mean
  - Standard errors differ, because standard regression assumes equal variance between the groups and uses a pooled estimate of variance
  - CI will therefore also be different
- Estimated slope: the difference between group sample means
  - The mean of a difference is difference of means
  - Inference is wrong because it does not account for dependent observations

62

Ex: LR w/ Robust SE (Wrong)

- In beta carotene datasets, a comparison of mean vitamin E levels at baseline and after 3 months of treatment (ignoring dose for the purpose of this illustration)
- Robust variance estimates: Compare
  - "`ttest vite0=vite1`", and
  - "`regress vite time, robust`" on reshaped data

63

Ex: Robust SE; No Clusters

```
. ttest vite0=vite1
Varbl | Obs    Mean    StErr.   StDev.           [95% CI]
vite0 |  45    8.025    .191    1.280    7.640    8.409
vite1 |  45    8.859    .135     .907    8.586    9.131
diff  |  45    -.834    .133     .894   -1.103    -.566
Ha: mn(diff) < 0   Ha: mn(diff) ~= 0   Ha: mn(diff) > 0
t =  -6.2598       t =  -6.2598        t =  -6.2598
P < t= 0.0000      P > |t|= 0.0000     P > t= 1.0000

. regress vite time if (time==0 | time==1) & ptid!=40,
       robust
                       Number of observations =    90
 vite | Coef.   StErr.       t     P>|t|     [95% CI]
 time |  .834   .234      3.57    0.001     .370  1.299
_cons | 8.025   .191     42.07    0.000    7.646  8.404
```

64

Correspondence

- Estimated intercept: the group 0 sample mean
  - Standard errors agree more here
  - CI differs a bit more, because of difference in degrees of freedom: t test used 44, regression used 88
- Estimated slope: the difference between group sample means
  - Inference is still wrong because it does not account for dependent observations
    - When no clusters specified, robust SE only accounted for possibility of unequal variances, not dependence of data

65

Example: LR with Clusters

- In beta carotene datasets, a comparison of mean vitamin E levels at baseline and after 3 months of treatment (ignoring dose for the purpose of this illustration)
- Robust variance estimates in clusters: Compare
  - "`ttest vite0=vite1`", and
  - "`regress vite time, robust cluster(id)`" on reshaped data

66

Ex: LR w/ Robust SE; Clusters

```
. ttest vite0=vite1
Varbl | Obs    Mean     StErr.   StDev.          [95% CI]
vite0 |  45    8.025     .191    1.280     7.640    8.409
vite1 |  45    8.859     .135     .907     8.586    9.131
diff  |  45    -.834     .133     .894    -1.103    -.566
Ha: mn(diff) < 0   Ha: mn(diff) ~= 0   Ha: mn(diff) > 0
t =  -6.2598       t =  -6.2598        t =  -6.2598
P < t= 0.0000      P > |t|= 0.0000     P > t= 1.0000


. regress vite time if (time==0 | time==1) & ptid!=40,
       robust cluster(ptid)
                       Number of observations =    90
 vite | Coef.    StErr.      t     P>|t|     [95% CI]
 time |  .834    .134     6.22    0.000     .564   1.104
_cons | 8.025    .192    41.83    0.000    7.638   8.411
```

67

Correspondence

- Estimated intercept: the group 0 sample mean
  - Standard errors agree more here
  - CI differs a bit more, due to degrees of freedom: t test used 44, regression used 88
- Estimated slope: the difference between group sample means
  - Standard error for slope is approximately the standard error for the mean difference
  - CI, t statistic, and P value are about the same but again are influenced by degrees of freedom

68

Effect of Correlated Data

- Note the problem when clustered data not identified
- The observations at baseline and 3 months were positively correlated
- Within clusters, the predictor of interest (time) differed
  - Each cluster had both a baseline and a 3 month measurement
- SE from classical regression to be too high
  - Conservative inference (loss of power)

69

Ex: Clustered Analysis; All Data

```
. ttest vite0=vite1
Varbl | Obs    Mean    StErr.   StDev.          [95% CI]
vite0 |  45    8.025    .191    1.280    7.640    8.409
vite1 |  45    8.859    .135     .907    8.586    9.131
diff  |  45    -.834    .133     .894   -1.103    -.566
Ha: mn(diff) < 0   Ha: mn(diff) ~= 0   Ha: mn(diff) > 0
t =  -6.2598       t =  -6.2598        t =  -6.2598
P < t= 0.0000      P > |t|= 0.0000     P > t= 1.0000


. regress vite time if (time==0 | time==1), robust
      cluster(ptid)
                       Number of observations =    91
 vite | Coef.   StErr.      t     P>|t|      [95% CI]
 time |  .798   .136     5.88    0.000     .525  1.072
_cons | 8.060   .191    42.20    0.000    7.676  8.445
```

70

Correspondence

- The patient with missing data at 3 months now influences the baseline mean
  - Thus we no longer have correspondences with the paired t test, which cannot use partial information on a subject
- Relative appropriateness of approaches depends on whether
  - “Missing at random”: Subjects missing data are similar to other subjects having the same values of the modeled variables
  - Interest in population differences or within subject differences

71

Aside: Handling Missing Data

- Two basic approaches
  - Omit cases with missing data
  - Impute missing data
    - Predict what missing data would have been

72

Aside: Omitting Cases

- If cases missing data are in some way different in (unknown) response values
  - Unbiased estimates of population that is not prone to missing data
    - Condition on such a population
  - Biased if cases missing data are different
- If cases missing data are "missing at random" based on modeled data
  - May lose power (relative to imputing)
    - Depends on amount of information derived from imputing

73

Aside: Imputing Data

- Predict missing data from a regression model or matching scheme
  - Can use all available data, including variables not otherwise included in the data analysis
  - Accurate to the extent you have the right prediction model
- Account for variability
  - Predict individual observation (mean + noise)
  - Multiple imputation for better SE

74

Correlated Data Analysis

- The robust variance estimates modify the standard errors, not the estimated slope
- Estimated slopes are the same as if every subject were independent
  - Each subject is not weighted equally if the sample sizes per subject are not the same
- This is OK for testing associations, because the standard errors will estimate the sampling variability correctly
  - But for esthetic reasons, we might prefer estimates which weight each subject appropriately

75

Correlated Data Analysis

- Regression parameter estimates will generally be "population" rather than "within individual" effects
- "Population effects"
  - Estimates refer to comparing different subjects who have different predictor values
  - "Marginal models", e.g., "Generalized Estimating Equations (GEE)
- "Within individual effects"
  - Estimates refer to comparing measurements on the same subject when he/she has different predictor values
  - "Random effects" in "Mixed models"

76

Weighted Regression

77

Weighted Analyses

- When sampling of individuals does not reflect their importance in the population, we should re-weight observation
- Biost 540 will address issues to choice of models with dependent outcome data more fully
- In this class, we will only consider
  - How to ensure accurate inference, and
  - How to obtain estimates which weight clusters equally in a simple manner

78

Choice of Weights

- The importance we should give each “cluster” (individual) should depend upon
  - the scientific question
  - the regression modeling of the predictors
  - the distribution of predictors across individuals
  - the reasons that we might have fewer observations for some subjects (e.g., nonignorable missingness?)

79

Weighted Regression

- Allows more emphasis to be placed on some observations than others
- Stata classification:
  - “Frequency weights”
  - “Analytic weights”
  - “Probability weights”
- In fact, all these weighting schemes actually use the same mathematical techniques
  - Vary in the way the results are presented

80

Weighted Regression in Stata

- Every regression command in Stata can take weights
  - `regress yvar xvar [wttype=expr] ...`
  - could substitute “logistic”, “poisson”, “stcox” for “regress”
  - square brackets [ ] are necessary
  - wttype is one of “fweight”, “pweight”, or (only for “regress”) “aweight”
  - expr is an expression typically involving some variable

81

Frequency Weights

- Adjust for duplicate cases that have the exact same measurements: Weights proportional to the frequency
- Sometimes for reasons of storage efficiency, we tabulate the number of cases having the same response and predictors
- We then have an additional variable that represents the frequency with which each observation occurs in the sample

82

Ex: Frequency Weights

- Frequency weights have greatest use in logistic regression
- For each combination of predictors have
  - One case counting number of subjects with response = 1
  - One case counting number of subjects with response = 0
- Often taken from registry or census data

83

Analytic Weights (Lin Reg Only)

- Adjust for cases which were measured with greater precision than others
- Weights proportional to the inverse of the variance
- E.g., the response is an average of several measurements
  - More measurements used in average = more precision
  - Number of measurements may not be the same for each case
  - Weights are the number of measurements used to compute the response

84

Uses of Analytic Weights

- Adjust for heteroscedasticity
- Analytic weights are inversely proportional to the variance of the individual case
  - We would need to know the variance in each predictor group
  - Then weight individuals according the value of their predictors
    - Can be quite difficult with multiple predictors
- (Sandwich estimator is much to be preferred)

85

Probability Weights

- Adjust “oversampling” or “undersampling” some elements of the population
- Sometimes some segments of the population have not been sampled in direct proportion to their importance in the population
- Weights proportional to the inverse of the sampling probability
  - E.g., multiple measurements on same individual
    - Each individual is equally important in the population
    - Each individual may not have the same number of measurements in the population

86

Use of Probability Weights

- It is the probability weights that we are most interested in for this course
- We can use them to adjust regression estimates in order to reflect equal emphasis placed on each individual
- We compute the “empirical probability” of sampling each observation as proportional to the count of observations for each “cluster”

87

Ex: Salary by Year 1990 - 1995

- How does geometric mean salary differ by year between 1990 and 1995
- Correlated observations on faculty
- Some individuals were hired between 1991 and 1995
- Fewer observations for those subjects

88

Ex: Classical LR (Wrong)

```
. drop if year <90
(11239 observations deleted)
. g logslry= log(salary)
. regress logslry year
                             Number of obs =    8553
  Source |   SS    df     MS F(  1,  8551) =  196.59
   Model |  17.4    1  17.4  Prob > F      =  0.0000
Residual | 757.3 8551  .089  R-squared     =  0.0225
   Total | 774.7 8552  .091  Adj R-squared =  0.0224
                             Root MSE      =  .29759
logslry |  Coef   SE      t     P>|t|     [95% C I]
   year | .0266 .00190  14.02   0.000  .023  .030
  _cons | 6.189 .17565  35.24   0.000  5.84  6.53
```

89

Ex: Only Robust SE (Wrong)

```
. regress logslry year, robust

Linear regression            Number of obs =    8553
                             F(  1,  8551) =  198.22
                             Prob > F      =  0.0000
                             R-squared     =  0.0225
                             Root MSE      =  .29759
        |        Robust
logslry |  Coef   SE       t   P>|t|    [95% C I]
   year | .0266 .00189  14.08 0.000  .023  .030
  _cons | 6.189 .17483  35.40 0.000  5.85  6.53
```

90

Ex: Cluster(ID) (Wrong Weights)

```
. regress logslry year, robust cluster(id)
Linear regression            Number of obs =    8553
                             F(  1,  1596) =  630.86
                             Prob > F      =  0.0000
                             R-squared     =  0.0225
Nbr of clusters (id)= 1597   Root MSE      =  .29759
        |       Robust
logslry |  Coef   SE     t   P>|t|     [95% C I]
   year | .0266 .0011 25.12 0.000   .025   .0287
  _cons | 6.189 .0989 62.57 0.000   6.00   6.383
```

91

Comments

- Identifying clusters made inference more precise
  - Estimates were the same
  - Positively correlated observations within clusters
  - Predictor (year) varied within clusters
- But weighting counted some individuals more than others
  - We might want to weight individuals equally

92

Ex: Cluster(ID); Weighted

```
. egen cnt= count(id), by(id)
. regress logslry year [pw=1/cnt], robust cluster(id)
(sum of wgt is   1.5970e+03)
Linear regression             Number of obs =    8553
                              F(  1,  1596) =   55.40
                              Prob > F      =  0.0000
                              R-squared     =  0.0077
Nbr of clusters (id)= 1597    Root MSE      =  .30503
        |       Robust
logslry |  Coef    SE     t    P>|t|     [95% C I]
   year | .0157 .0021  7.44  0.000  .0116  .0198
  _cons | 7.179 .1946 36.88  0.000  6.797  7.561
```

93

Comments

- This weighted individuals equally
- However, because the individuals with less data were newer hires, there is some bias
  - Older hires apply to all years, newer hires to recent years
  - We would really rather have a model that considered the difference in salaries in order to get something closer to the within individual
  - Population effects are probably less of interest
    - Consider within individual effects by adjusting for year 95 data

94

Comments

```
. egen lslry95=max(logslry), by(id)
. regress logslry year lslry95 [pw=1/cnt] if year <95, robust
   cluster(id)
(sum of wgt is   1.2309e+03)
Linear regression             Number of obs =    6956
                              F(  2,  1522) =15908.31
                              Prob > F      =  0.0000
                              R-squared     =  0.9433
Nbr of clusters (id)= 1523    Root MSE      =  .07172
        |        Robust
logslry |  Coef     SE       t   P>|t|   [95% C I]
   year | .03661 .00057  64.16 0.000 .0355 .0377
lslry95 | .98917 .00595 166.17 0.000 .9775 1.001
  _cons | -3.385 .07404 -45.72 0.000 -3.53 -3.24
```

95

Comments

- Averaging over the past five years, but counting each faculty member equally:
  - Faculty have averaged 3.7% raises (95% CI 3.6% to 3.8%) over the past five years
  - Highly statistically significant (P < 0.0005)
- (Note that we had to decide whether we should weight years equally or faculty equally)

96