Epidemiology Methods III Notes

Class 1: Introduction to applied epidemiologic methods: Linear regression (part 1)

Purpose of different epidemiologic studies

Cross-sectional studies: Hypothesis generation (Prevalence- Incidence Bias: we cannot use x-sectional studies for causality, because we are looking at a snapshot in time for our sample and dealing with prevalent outcomes, whereas for causality we need incident outcomes)
Case Control Studies: Introduces Temporality
Cohort Studies: Causality Investigation

Statistical Inference in Epidemiology

- We are going to examine the associations between independent variables (our exposures) and dependent variables (our outcomes) to determine if an association exists in the population. Our goal is to determine if we can predict an outcome knowing the exposure, based on the association.

- There are two potential sources of error that can be found in our computed measures of association

    - Bias (systematic errors): affects the accuracy (point estimate, or how well we measure the association) and can be due to things such as confounding or information bias

    - Random variation (random error): affects the precision (confidence interval around the point estimate) and can be due to sampling errors

    - Statistical inference (P-values and confidence intervals) is the set of tools that we use to deal with the “sampling uncertainty “ after accounting for all known sources of bias.

Population	Theoretical and non-enumerable set that is the target of scientific inference	**Parameter**	Measure of interest (Mean, RD, OR) in the theoretical population	**Standard error**	Quantitative indicator of the spread of sampling distribution
Sample	Study sample, real subset of the theoretical population	**Parameter estimate**	Measure of interest (Mean, RD, OR) in the real sample	**Standard Deviation**	square root of sample variance: used to estimate the standard error (SE = SD/√n)

We go from Variance (sample) > SD(sample) = √Variance > SE of the mean (Sample) = SD/√n. The standard error of the mean is a quantitative indicator of the true population variance, based on our sample.

How to use statistical inference to ensure the proper sampling distribution of a measure of association:

I. Statistical Hypothesis Testing of the null hypothesis: mean difference = 0

When we use the assumption that our sampling distribution is normal (Gaussian), we can use the following formula to test the null hypothesis that there is no association between exposure and outcome (Ho: X = Xo, or X -Xo = 0):

Test statistic = X – Xo / SE (usually Z or T statistic, can also be extended to X^2 or F statistic)

We can use the test statistic and find the corresponding P-value to compare it to α:

- P value vs α of 0.05 for a two tailed test (2.5% at each end of the Gaussian curve) the cut off for 95% (1-0.05) for a t test is 1.96. If our P-value is smaller than 0.05 (aka our test statistic is greater than 1.96, we can reject the null hypothesis)

- α = 0.05 is our type I error (this represents false positives, where we find an association and reject Ho (no association) when Ho is true and there is no association)

- βis our Type II error (this represents false negatives, where we do not find and association and accept Ho (no association) when Ho is false and there is an association). 1 – β represents our power in our statistical analysis

II. Estimation: a 95% confidence interval

We can also use the standard error to construct a confidence interval around our parameter estimate:

95% CI = Parameter estimate +/- 1.96 * SE

- If our CI does not contain 0, our test statistic is significant and we can reject the null Hypothesis

The Scientific Question

Etiologic model in epidemiologic studies: 1 exposure -> 1 Outcome
Prediction models in epidemiologic studies: Multiple exposures -> 1 Outcome

The typical question in etiologic epidemiology is about a causal connection between exposure and disease. The variables can be dichotomous, categorical or continuous, and the data analysis will examine the association in two ways:

- Is there an association?
- What is the magnitude of the association?

Analyzing an association using a t-test (not linear regression): A Ttest is a special application of the linear regression with categorical predictors (Comparing the means of an outcome between two groups).

```
PROC TTEST;
        CLASS (X); /*Categorical Predictor (AKA your groups)*/
        VAR (Y); /*Continuous or Categorical Outcome*/
        Run;
```

- The statistics section will give you the mean (X), standard deviation and n for each groups. These can be used to find the standard error (SE = SD/√n) and the T-statistic (ΔX/ SE) and 95% CI (Parameter estimate +/- 1.96 * SE).
- However, the statistic section will also give you the standard error, T-statistic, P-value and 95% CI.
- If the test statistic is > 1.96 or the P-value is < 0.05, the test is statistically significant. If the CI does not contain 0, the test is also statistically significant

Simple Linear Regression Model

The dependent Variable (Y) is continuous, and the Independent Variable (X) can be continuous or categorical (this is the case for a 2 sample t-test, which is a particular application of SLR). We specify the mathematical form for the relationship between the two variables

Y (predicted) = **β0 + β1 *X (this is the equation of the fitted regression line, therefore there is no error component)**

Y (observed) = **β0 + β1 *X + Error = Y (predicted) + Error (there is an error term because of the random differences between the expected and observed values, or residuals)**

- β0 is the **intercept**, and it is the predicted value of Y when X = 0
- β1 is the **slope** and it is the predicted difference in Y for a 1 unit increment in X. This value quantifies the effect of X on Y.

Analyzing the association using simple linear regression

```
PROC GLM (or PROC REG, but PROC GLM is more advanced)
        MODEL Y=X;
        Run;
```

SAS Output

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square (SS/Df)	F value	P Value
Model	K-1	SS (Group)	MSG = SS/ (K-1)	MSG/MSE	
Error	N-K	SS (Error)	MSE = SE/ (N-K)		
Total	N-1	SS total			

The F-value is a test statistic, similar to the T-test. Based on the table above, F-test = MS regression (MSG) / MS residual (MSE) = [(SSY – SSE) / regression df] / [SSE / residual df]

SSY = (Yi – Y)^2
SSE = (Yi – Y^)^2
R-squared = (SSY-SSE)/(SSY)

The R-Square value represents the percentage of variance in Y that can be attributed to X. The closer it is to 1 (aka SSE is closer to 0), the better it is.

Analyzing the association with confounding (A is the potential confounding variable)

```
PROC GLM
        MODEL Y=X A;
        Run;
```

SAS output

Type I SS: adjusts the variables sequentially and Type III SS: adjusts the variables simultaneously. Type III SS is the SS that we always want to use. When looking at the SAS output for adjusted variables, a significant p-value (<0.05) means that the predictor variable is an **independent predictor** of the outcome, (Because it is still significant after adjusting for other variables). Like wise, a non-statistically significant p-value means that the variable is not an independent predictor of the outcome.

Properties of a confounding variable (Source of bias)

- Must be a cause of disease or at least a marker (surrogate) of an actual cause of the disease
- Must be distributed differently in exposed and unexposed
- Cannot be an intermediate step in the causal pathway between exposure and disease

Types of confounders

- Known (can be measured accurately)
- Known (measured with error)
- Known (not measured)
- Unknown

10% confounding rule: [β1 (crude) - β1 (adjusted)] / β1 (crude)

- If the difference is greater than 10%, we can assume that the variable A is a confounder
- If β1 (crude) > β1 (adjusted), it is **positively confounded** (the crude estimate is biased away from the null)
- If β1 (crude) < β1 (adjusted), it is **negatively confounded** (the crude estimate is biased towards the null
- **The 10% rule is qualitative measure, not a statistical one**

Math Review ☺

- The value of x such as that 10^x = y is call the logarithm of y with the base 10 and is written log 10 (y)
- A special base is log e (y) where e=2.718 (logarithm to the base e is called the natural logarithm)
- For a very small x, e^x = 1+x and log e (1+x) = x

Arithemtic of exponentials

- **e^a * e^b = e^(a+b)**
- **e^a / e^b = e^(a-b)**
- **e^0 = 1**
- **e^-a = 1/e^a**

Arithmetic of logarithms

- **Log (a*b) = log (a) + log (b)**
- **Log (a/b) = log (a) – Log (b)**
- **Log (a^x) = x * log (a)**
- **Log (1) = 0**

Class 2: Linear Regression

Review of interaction/effect modification:

Interaction is present if

- Observed joint effect (of A and B) seen in data =/= Expected joint effect (of A and B) as predicted by model
- The effect of A (on disease outcome) varies across the strata of B and vice versa.

Both definitions requires us to specify the measure of association to represent the causal relation

- If our measure of association is a **rate or risk difference**, the interaction is said to be on the **additive scale**
- If our measure of association is a **rate or risk ratio**, the interaction is said to be on the **multiplicative scale**

SAS code

Step 1

```
Proc Sort;
by B;

Proc GLM;
Model Y = A;
By B; /* stratify by B in order to see if the effect of A differ across strata (suspicion of effect modification)*/
```

Step 2

```
Proc GLM;
Model Y= A B A*B; /* the term A*B is formally testing for interaction*/
```

Step 3

```
Proc GLM;
Model Y = A B; /*add B to the model as a potential confounder*/
```

SAS Output

Is the interaction is not statistically significant, it means that the association between A and Y does not differ across level of B. However B could still be a confounder (10% rule), and an independent predictor (significant p-value).

- When there is no interaction between 2 predictors, the GLM slopes will be parallel
- If there is interaction between two predictors, the GLM slopes will not be parallel

Class 3: Cross Sectional Studies

The simplest way to show the relationship between 2 binary variables is a 2 x 2 table

		Disease		
		Yes	No	
Exposure	Yes	a	b	a+b
	No	c	d	c+d
		a+c	b+d	Total

- For Cohort studies, the risk ratio is the prevalence ratio: [a/a+b]/[c/c+d]
- For Cohort studies, the risk difference gives us the attributable risk due to exposure and is: [a/a+b] - [c/c+d]
- For Cohort Studies, the Odds ratio is the ratio of 2 odds and can be expressed in 2 ways
  - Disease odds (cohort): Disease odds among exposed / Disease odds among unexposed = [a/b]/[c/d]
  - Exposure Odds (case control): Exposure odds among diseased / Exposure odds among non-diseased = [a/c]/[b/d]
  - Disease odds = Exposure odds = (a*d)/(b*c)
  - **When the disease is rare (a is low and c is low) the risk ratio ~ odds ratio**

If there is no association between the exposure and disease, then the disease prevalence will be the same in the exposed and unexposed: n1/T = a/m1 = c/m0: **The null value for RR and OR is 1, and the null value for RD = 0**

How to interpret Measures of Associations:

- EOR: People with D have xtimes higher odds of having E than people without D
- DOR: People with E have xtimes higher odds of having D than people without E
- RR: People with E have a xtime greater risk of having D than people without E
- RD: x% of D can be explained with E

Statistical Hypothesis testing for a 2x2 table

Step 1: Formulate the null hypothesis
Ho: No association (OR=1, RR=1,RD=0)

- Calculate the probability using a chi square test
- Use the P-value for the test to accept or reject the null

Step 2: Find the test statistics

Pearson's Chi-square: Easy to use, but doesn't work with thin data.
X^2 = [observed – expected]^2/expected
X^2 = [(ad-bc)^2)*T]/[n1n0m1m0]
Degrees of freedom = (row -1)(column – 1)

- Do not use if there is less than 5 observations per cell
- For a large sample, Pearson's Chi-square and Fisher's exact test will yield the same p-value

Continuity-corrected chi-square: Chi-square alternative used for small sample sizes

Fisher's exact test : Calculated as a one tailed, but you can't just multiply the result by 2 for the two tailed result since the upper and lower tail values are different; It is always applicable, even with small cell counts

Logit Chi-square (Woolf): X^2 used on the log scale because OR is not normaly distributed (i.e, 2 and ½ are at different distance for the null which is 1) but Ln(OR) is normally distributed (Ln(1) = 0 which is the null on the log scale)

- Z = X-Xo/SE
- OR = e^B therefore, B = Ln(OR)
- Z = ln(OR) – Ln (1) / SE[ln(OR)], and **(Ln(1) = 0)**
- X^2 ~ Z^2 = [ln(OR)]^2/ SE[ln(OR)]^2
- X^2 = [ln(OR)]^2/ Var [ln(OR)] and **(SE^2 = Variance)**
- Var [ln(OR)] = 1/a +1/b + 1/c + 1/d

Miettinen Confidence Intervals: Applicable to all Chi-squares above solving for SE.

- X^2 = [ln(OR)]^2/ Var [ln(OR)]
- X^2 = [ln(OR)]^2/ SE[ln(OR)]^2
- SE[ln(OR)] = ln(OR)/√X^2

95% Ci for the ln(OR): ln(OR) +/- 1.96*SE[ln(OR)] = ln(OR) +/- 1.96 * ln(OR)/√X^2

95% Ci for the OR: e^[ln(OR) +/- 1.96 * ln(OR)/√X^2]

Particular application: Logit Method (Woolf)
Var [ln(OR)] = 1/a +1/b + 1/c + 1/d
So SE[ln(OR)] = √[1/a +1/b + 1/c + 1/d]

95% CI for the ln(OR): ln(OR) +/- 1.96*SE[ln(OR)] = ln(OR) +/- 1.96 * √[1/a +1/b + 1/c + 1/d]

95% CI for the OR: e^(ln(OR) +/- 1.96*SE[ln(OR)])

SAS code for 2 *2 tables

```
Proc freq;
Tables x*y /chisq chm1;
/*CMH is the Cochran mantel haentzel for adjusted analysis*/
```

SAS output

- You are interested in the percents for the X variables (so if X is in row, you want row percents)
- Under relative risks, the case control CI is the logit/woolf CI (this is why the parameter is not in the middle of the confidence interval

SAS Estimates of the Relative Risk

- Case-control (Odds ratio): This is the disease odds ratio (cohort) or exposure odds ratio (case control) or prevalence odds ratio (cross-sectional).
- Cohort 1 (Column 1 risk): this is the disease risk ratio (a/a+b)
- 95% confidence bounds (asymptotic): These are the logit based confidence intervals (Woolf)
- 95% confidence bounds (exact): on the regular scale, the log transformed parameter estimate may not be in the exact middle

SAS Estimates of the Common Relative Risk: This is used for pooling estimates of a series of 2x2 tables (it will be the mantel-haenszel estimates)

SAS code for 2*n tables

```
Proc freq order = formatted;
Tables X*Y / chisq cmh1 nocol nopercent;
Format X quartile.;
```

Degrees of freedom = (row-1)(column-1)
When we have more than a 2*2 table, we cannot use the Pearson chi-square

Chi-square test of the null hypothesis of a 2*n table with (row-1)(column-1) degrees of freedom: The prevalence of disease (Y) is the same across all levels of exposures (X). Aka, there is no association between exposure and disease

The CMH (Cochran Mantel Haenszel or Mantel Haenszel Pooled Chi-square statistic) **is the 1-degree of freedom chi-square is mantel's test for trend:** it is a test for the following null hypothesis: there is no linear component to the trend across the X quartile. You perform the test by assigning an X value on an ordinal scale to each parameter. However, the test for linear trend is not a test for monotonic dose-response, but a test for monotonic trend (in the same direction)

CLASS 4: CROSS-SECTIONAL STUDIES TABULAR METHODS-ADJUSTMENT FOR CONFOUNDING

Regression Data will give you the expected values, so the OR between the strata will be the same

Observational Data will give you observed values, so the OR between strata will slightly differ due to SRS

Counterfactual: Being able to examine the relationship between E and D if E never happened (that way all other factors would be the same and we would not have to worry about confounding). So observe the E group until D happens, then go back in time and reexamine the same people as unexposed.

Tabular Methods for assessment and control of confounders: For confounding, we use the etiologic model and assume that 1 exposure -> 1 disease

1. **Stratify the data on the confounder (create a 2x2 table of E and D for each stratum of the confounder)**
2. **Compute a measure of association between E and D for each stratum of exposure**
3. **Decide if it is acceptable to ignore any differences between the stratum specific measures of association (test for interaction or effect modification)**
    a. **Using the Breslow-day test:**
        - X^2 = Σ (observed – expected)^2/Σ variance for each strata
        - Parameters are k-1 df, and p-value as usual
        - **If the p-value is statistically significant, then the stratum specific OR differ significantly and we should not report a summary OR.**

    b. **Logit test (Woolf)**
        - Wi = 1 / [1/a + 1/b + 1/c + 1/d] = 1/var [ln(OR)]
        - X^2 = Σ [ln(ORi)^2 - ln(ORm-h)^2] / var [ln(ORi)] = Σ [ln(ORi)^2 - ln(ORm-h)^2] * Wi
        - Parameters are k-1 df, p-value as usual

4. **If strata 1 ~ strata 2 then compute a pooled estimate (e.g. adjusted OR)**
    a. **Method 1: Using mantel-haenszel summary estimators**

		Disease		
		Yes	No	
Exposure	Yes	a	b	m1
	No	c	d	m0
		n1	n0	Total

Let RRi be the stratum specific measure of association and let Estimator m-h be the pooled measure of association

Estimator = Σ(Wi*RRi)/ Σ (Wi)

- For Cumulative Incidence, Prevalence or Risk Ratio: RRmh = Σ (Wi*RRi)/ Σ (Wi) = Σ (c*m1/T)*[(a/m1)/(c/m0)]/ Σ (c*m1/T) = Σ (a*mo/T)/ Σ (c*m1/T)
- For Odds Ratio: ORmh = Σ (Wi*RRi)/ Σ (Wi) = Σ (b*c/T)*[(a/b)/(c/d)]/ Σ (b*c/T) = Σ [(a*d)/T]/ Σ [(b*c)/T]

Assumptions: No residual confounding within strata because the strata is homogenized (the effect of exposure on the disease is the same within all strata because we assume no interaction between E and the confounders)

b. using the Logit-based method

Ln(OR) = Σ (Wi*ln(OR))/ Σ (Wi)
OR = e^ [Σ Wi*ln(OR))/ Σ (Wi)]
Wi = 1 / [1/a + 1/b + 1/c + 1/d] = 1/var [ln(OR)]

Assumptions: same as the MH estimator, but not good if the data is thin or any cell has 0, therefore the Mantel-Haenszel method is the method of choice!

5. **Compare the pooled estimate to the crude estimate**
   a. Use the 10% rule of thumb: |(lnORadj –lnORcrude)|/(lnORcrude) or (crude – adjusted)/(crude) to find out the sign of confounding
   b. If the adjusted differs from the crude by more than 10%, present the adjusted estimate
   c. Be careful about using this is the crude ratio is close to 1 (the null) or if you have a small sample size.

Hypothesis testing and CI for adjusted OR
Null hypothesis testing Ho: ORmh = 1

		Disease		
		Yes	No	
Exposure	Yes	a	b	m1
	No	c	d	m0
		n1	n0	Total

Expected value = n1i*m1i/Ti
Variance = n1in0im1im0i/Ti^3
X^2 = Σ (observed – expected)^2/Σ variance, p-value as usual

95% Ci for the ln(OR): ln(OR) +/- 1.96*SE[ln(OR)] = ln(OR) +/- 1.96 * ln(OR)/√X^2
95% Ci for the OR: e^[ln(OR) +/- 1.96 * ln(OR)/√X^2] = OR*e^(1+/-1.96*1)/√X^2)

Logit-Based Estimator (Woolf)

Wi = 1 / [1/a + 1/b + 1/c + 1/d] = 1/var [ln(OR)]

X^2 = ln(OR)^2/ var [ln(OR)] = ln(OR)^2* Wi

95% CI for Logit Based Estimator: e^(ln(OR) +/- 1.96*SE[ln(OR)])

SAS code for tabular assessment of confounders

```
Proc freq order = formatted;
Tables C*E*D / CMH1 NOCOL NOPERCENT;
```

- The case control OR is the adjusted value that we want to use and compare with the crude OR
- The CMH (Cochran Mantel Haenszel or Mantel Haenszel Pooled Chi-square statistic) is testing Ho: OR =1 to make sure that there is correlation between exposure and disease is the Mantel's Test for trend.
- The Breslow-Day Test for homogeneity is a test for interaction/effect modification: It tests to make sure that the OR are similar across each strata of the confounder (i.e. there is no effect modification so if the p-value is significant, that means that the s strata 1=/= strata 2 therefore there may be effect modification and we should not pool the data

In Summary:

- **If strata 1 =/= strata 2 we have potential effect modification -> report stratum specific estimates.**
- **If strata 1 ~ strata 2 , we can pool strata 1 and strata 2. If crude =/= pooled , then we have potential confounding ->report adjusted estimate**

CLASS 5: Unconditional logistic Regression Basics

Undesirable features of the Simple linear regression model:

- The Y is usually constrained between 0 and 1 (ie yes vs. no)
- The error around the expected vale does not have a normal distribution
- The coefficient B1 is the <u>predicted</u> difference in the probability of Y=1 for a 1 unit change in X. It is a risk difference rather than a risk ratio

Change the probability scale from 0 / 1 to 0 / infinity by using log scale to fix the binomial distribution: Odds = P/(1-P)

Odds (Y=1) = (Pr Y=1)/(1- Pr Y=1) = Bo + B1*X
Odds = P/(1-P) = Bo + B1*X

Change the probability scale from 0/infinity to –infinity / +infinity: log odds = Ln [P/(1-P)]

Ln (P)/(1- P) = Bo + B1*X
Log (risk of disease/risk of no disease) = Bo + B1*X

Given that ln (a/b) = ln (a) – ln (b)

Ln (P) – Ln (1- P) = Bo + B1*X
Ln (P) = Bo + B1*X + Ln (1- P)
Ln (P) = Bo + B1*X + Ln [(1- Bo + B1*X)]

And

e^ [Ln (P)] = e^[Bo + B1*X + Ln (1- Bo + B1*X)]
P = e^(Bo + B1*X) / [1+ e^(Bo + B1*X)]

The really easy way to remember that is that

- **Probability = odds / 1+ odds**
- **Odds (of event) = probability (of event) / probability (of no event) = probability (of event)/1 – probability (of event)**

Logistic regression

- The model yields odd ratios as results
- It allows for sequential or simultaneous assessment of multiple risk factors, confounders and interaction

SAS code for logistic regression

```
Proc logistic descending; /* used to make sure that it is sorted in 1, 0 in the tables for the outcomes*/
Model Y = X /RL; /*this is to make sure that we get the 95% CI (or risk limits) */
```

B1 interpretation (parameter estimate for X):

- Ln(OR) = Bo + B1*X, thus OR = e^B and B = Ln(OR)
- [log odds Y =1|X =1]/ [log odds Y =1|X =0]
- [log odds disease | exposed]/ [log odds disease | unexposed]
- log odds in disease for 1 unit increment in X

Y intercept interpretation

- Pr (D|no exposure) because X = 0 (categorically that represents the unexposed)
- E^Bo is the odds of disease in the reference group

95% CI for the Ln(OR) = B1 +/- 1.96*Se(B1)
95% CI for the OR = e^[B1 +/- 1.96*Se(B1)]

Wald Chi-square (Chi square in Logistic regression):
The null hypothesis is Ho: B1 =0 or Ho: OR =1
Z = B1/SE(B1)
X^2 ~ [B1/SE(B1)]^2 df = 1

The Delta Method

	Intercept	X
Person A	B0	a
Person B	B0	b
Delta	0	(a-b)

Delta = (a-b)
Log (odds |X=a) = Bo + B1*a
Log (odds |X=b) = Bo + B1*b
Log (odds |X=a) - Log (odds |X=b) = Bo + B1*a - (Bo + B1*b) = B1(a-b)
OR (a vs. b) = e^[B1(a-b)]

B1 = Delta [log odds (Y=1)]/ Delta X = Log OR per 1 unit increment in X

By default the 1-unit increment delta is 1, but for some biological measurements, we set delta = 1 SD.

Confidence Intervals for any delta exposure (Usually Delta = 1 so we omitted it in the past)

95% CI : e^[delta*B1 +/-1.96*delta*SE (B1)]

Conclusion

- The OR can be expressed as an exponential function of X (e^B1)
- The Log(OR) can be expressed as a linear function of X (B1)

CLASS 7: Unconditional Logistic Regression Indicator Variables

- Binary Variable: Takes two values
- Nominal Variable: takes a discrete number of values, without any assumptions concerning order or distance
- Ordinal Variable: takes a discrete number of values, which are rank ordered. The distance between values may not be constant
- Categorical variable: Can be binary, nominal or ordinal
- Continuous variable: Takes an infinite number of value on an arithmetic scale, but the value are rank order and the distance between values is equal throughout the scale.

Modeling continuous variables in logistic regression

B1 = Delta [log odds (Y=1)]/ Delta X = Log OR per 1 unit increment in X

OR = e^B1*Delta X = odds ratio associated with a one unit increment in X.

We go from ratio to difference because log (a/b) ⇔ log (a) – log (b)

Modeling categorical variables in logistic regression

- If you use regular numbers to indicate your categories, SAS will treat the quartile as a continuous variable. Therefore, you should use indicator variables (0, 1) to force SAS to consider them as a categorical variable.

- Use indicator variables: assign 0 to the reference group, and 1 to the exposed group

- Pick the natural low risk or unexposed group as your reference level, and if this is not an obvious choice, pick the largest group to have the most stable statistical analysis

Example: Delta Method (with indicator variables)

Original	EX	SMK
Never Smoker (0)	0	0
Former Smoker (1)	1	0
Current Smoker (2)	0	1
B*Delta	B1*Delta	B2*Delta

- Log odds (Y=1) = Bo + B1 EX + B2 SMK

- Bo is the predicted log odds when EX=0 and SMK = 0: Thus e^Bo is the predicted log odds of disease in the reference group

- B1 is the predicted log odds ratio when EX = 1 (Versus EX =0): Thus e^B1 is the predicted odds ratio for former vs never smokers

- B2 is the predicted log odds ratio when SMK = 1 (Versus SMK =0): Thus e^B2 is the predicted odds ratio for current vs never smokers

For current vs former smokers

Original	EX	SMK
Current Smoker (2)	0	1
Former Smoker (1)	1	0
Difference	-1	1
B*Delta	B1*Delta	B2*Delta

OR = e^(B1*-1 +B2*1) = e^(B2-B1) = e(B2)/e(B1)

Interpreting SAS output for proc logistic with an X variable subdivided by indicator variables

X has categories labeled 0, 1, 2, 3 (we will make X = 0 the reference level)
If X = 1 then X1=1; else X1=0;
If X = 2 then X2=1; else X2=0;
If X = 3 then X3=1; else X3=0;

Proc logistic;
Model Y = X1 X2 X3 / RL; /* this gives you the confidence intervals */

Interpreting the output

Intercept = Bo = Odds of disease in reference level so e^Bo = Pr (Odds/1+Odds)

Parameter estimates (B1) for X1, X2 and X3 represent the slope: This is the difference in log odds of D for the Quartile Xi compared to the reference level.

Odds Ratio are the e^B1 and estimate the odds ratio of D in Xi when compared to the reference level

*** Test for linearity: There is not a 1 degree test for trend (CMH) in proc logistic, but comparing the difference between B1 of each quartile is a qualitative estimate of the linearity assumption: This is because linarity implies that a 1 unit distance between 2 variables should be the same regardless of the position on the line, so if the values are the same then you can assume that the linearity clause holds.

How to calculate the OR in 2 different categories other than the reference level

Original	(B1) X1	(B2) X2	(B3) X3
Person A (Upper Category, here X3)	0	0	1
Person B (Lower Category, Here X1)	1	0	0
Difference	-1	0	1
B*Delta	B1*(-1)	B2*(0)	B3*(1)

Log Odds (Y) = B1*(-1)+ B2*(0)+ B3*(1)
OR (X3 vs X1) = exp (B3-B1) = exp (B3) / exp (B1) = OR 3/ OR 1

How to test if two level of a dummy coded (or indicator) variable are different from one another:

Method 1: test statistic by Hand

The OR to compare B1 vs B2 (Ho: OR (B1 vs B2) =1 or Ho: B1 – B2 = 0) is

OR (B1 vs B2) = Exp (B1 – B2) = exp (B1) / exp (B2)

Z = B1 – B2 / Se (B1-B2)

X^2 = (B1 – B2)^2 / Var (B1-B2)

Since Se (B1 – B2) =/= Se(B1) – Se(B2)

Var (B1-B2) = Var (B1) + Var (B2) – 2* Covar (B1,B2)

*** use the / COVB option in the model statement to get the covariance

- Use Se (B1) ^ 2 = Var (B1)
- Use Se (B2) ^ 2 = Var (B2)
- Get the Cov (B1, B2) from the /COVB option
- Find Var (B1 – B2) then Se (B1 – B2)
- Find Z then X ^2 and p-value

Method 2: Recoding

The reference level of the indicator variable is the level for which all Xi = 0
Recode the lower category as all Xi = 0, then run the regular proc logistic

Class 8: Additional Computations and Confounders in Logistic Regression

Computing a Ci for the OR using logistic regression output

Single Variable vs Reference

OR = exp (Delta * Bi)
95% Ci : exp [(Delta*Bi) +/- 1.96 Se (Delta*Bi)]
95% Ci : exp [log OR +/- 1.96 Se (log OR)]

Multiple Variables

OR = exp [(Delta * Bi) + (Delta * Bj)]
95% Ci : exp {[Delta * Bi) + (Delta * Bj)]+/- 1.96 Se [(Delta * Bi) + (Delta * Bj)]

Compute Se [(Delta * Bi) + (Delta * Bj)] from the output

Var [(Delta * Bi) + (Delta * Bj)] = Delta ^2 Var (Bi) + Delta ^ 2 Var (Bj) + 2*Delta (i)*Delta (j)*Cov(Bi,Bj)
Se [(Delta * Bi) + (Delta * Bj)] = Sqrt {Var [(Delta * Bi) + (Delta * Bj)]}

How to assess confounding in logistic regressions

Tabular methods: Ln (Crude) – Ln (Adjusted) / Ln (Crude); need to be on the log scale

Logistic method: B crude – B adjusted / B crude; no need to transform since we are already on the log scale

SAS Code

```
Proc logistic;
Model Y = X / RL; /*Crude estimate*/

Proc Logistic;
Model Y = X A / RL; /* Adjusted estimate:
This is the main effects model since we do not have any interaction term*/
```

To evaluate confounding in proc logistic after adjustment, you need to use the parameter estimate: they are on the log scale, which insures linearity (one advantage of log scale)

Residual Confounding

When you model continuous variables as categorical variables, our adjusted estimates may not be too different from your crude estimates. However you are losing some data quality, and you may have very different adjusted estimates: this means that making the variables categorical reduced the impact of confounding, so we had residual confounding using arbitrarily categorical variables.

Summary

- Confounders are easily included in logistic regression as additional independent variables
- Confounding in logistic regression is assessed just as it was in tabular methods, by comparing the unadjusted parameter estimate to the adjusted parameter estimate
- Having indicator variables for our categorical variables allows us not to be constrained into the shape/data relationship

CLASS 9: Unconditional Logistic Regression Statistical Hypothesis Testing

Wald Chi Square: (B/se)^2

This is used as our test statistic for null hypothesis when we are looking at a single coefficient (Ho: B1 = 0)

The Test Statement

This is used to simultaneously test multiple coefficients (Ho: B1 = B2 = B3 =0)

SAS Code

```
Proc logistic;
Model Y = X;
NULLTEST: Test X = 0;
```

/* If X is the only variable for the test statement, the Null test chi-square will be the same as the Wald chi-square (AkA there is no association between levels of X and Y)*/

/*SAS will give you an intercept only fit (B0) and an intercept with covariates (B0 + BiX). The difference between the 2 would give you the Chi-Square associated with the Bi: Since there are no variables in the model (The B0 would cancel each other in the difference)*/

```
Proc logistic;
Model Y = X1 X2;
NULLTEST: Test X1 = X2 = 0;
```

We are now interested in testing the following null Hypothesis (Ho: B1 = B2 = 0) The null hypothesis implies that none are significant, and this would mean that all the variables are not significant, so we are testing for who has the D in our sample, not why they have it (E); Odds -> Pr (Odds/ 1+ Odds). This model will have (r-1)(c-1) df because there are multiple variables in the model.

If you use this for different variables, the null hypothesis implies that X1 and X2 are not associated with D after adjustment for each other. This is not very useful in etiologic epi (1 exposure -> 1 outcome) but is useful when we are looking at 1 variable with multiple levels (with indicator variables for each). The null hypothesis would then be stated as: The prevalence (odds) of D is identical in all levels of X.

The Likelihood Ratio Chi-Square

Regression models force a strict mathematical relation between the independent variables and the outcome. The Model fit using the likelihood ratio tells us how well our predicted regression line fits the data: in other words, we figure out what is the likelihood of the model, given the data. (Similar to R^2 in linear regression)

The log likelihood: the key piece of information is the -2LoG L in the proc logistic output. Likelihood = probability of observing Y = probability of each Yi, since they are all independent. How do you find your probability of Y? The 2 outcomes are 0 and 1 for probabilities

- Log odds (Y=1) = B0 + B1X
- Odds (Y=1) = e^(B0 + B1X)
- P = Odds / [1+ Odds] = e^(B0 + B1X) / [1 + e^(B0 + B1X)]
- Pr (Y=0) = 1- Pr (Y=1)

Likelihood is a probability: If Probabilities are between 0 and 1, Log (Probability) are between - & and 0. –log (Probability) are between 0 and +&, and we use -2Log (probability) for statistical testing reasons. In summary, the -2log only takes positive values and since 1 is the highest probability, the closer to 0 our -2log is, the better it is.

The Log-likelihood is used for hierarchical models. This occurs when one model as all of the variables of the other and more, plus both models use the same number of observations (on the same sample)

For example
Reduced model = Model 1: B0 + B1X
Full Model = Model 2: B0 + B1X + B2Z

Model 2 is like an “extension” of Model 1. And we examine the improvement of fit by comparing them. If the full model is no better than the reduced model for fit, then the likelihood (-2log) of the full model should be close to that of the reduced model. This is testing the following null hypothesis Ho: there is no improvement of fit by adding new variables to the reduced model <-> all of the additional coefficient estimated by the full model = 0.

This Ho is tested with the likelihood chi-square test
X^2 = -2Log reduced –(-2log full) = 2logfull = 2 log reduced with df = df full – df reduced (also called the likelihood ratio test because log a – log b <-> Log a/b)

Summary: The likelihood ratio test is concerned with the association of specific predictors in the model, but is not an assessment of confounding!

CLASS 10: Unconditional Logistic Regression Interactions

Logistic associations can be show on an additive or multiplicative scale:

Let's take Y = Bo + B1X + B2Z
The B1 and B2 represent the difference in log odds

Log Odds = B1 + B2 (additive scale using the delta method)
OR = e^B1 * e*B2 (multiplicative scale when we convert from log odds to odds ratio)

Linear regression is testing for an additive interaction:

Observed joint effect of two risk factors =/= Expected joint effect of these two risk factors

Logistic regression can test for both an additive or multiplicative interaction

Predicted joint effect (not observed because the regression derived OR are predictions) =/= expected joint effects of these two risk factors

- Both tests can also be interpreted as: "The effect of one risk factor on the disease does not differ across the levels of the second risk factor (no effect modification)"
- If there is no effect modification, you can present the estimates from your main effect model (independent predictors)
- Evaluating interaction in Tabular methods: Use the breslow-day test for homogeneity of the odds ratios in proc freq.
- Evaluating interaction in logistic regression: you need to make 2 models, and the second one will have an interaction term.

"Main effect model"
Proc logistic data = a;
Model outcome = exposure 1 exposure 2
Log odds (outcome) = B0 + B1 exposure 1 + B2 exposure 2

"Interaction model"
Proc logistic data = a;
Model outcome = exposure 1 exposure 2 exposure 1 * exposure 2
Log odds (outcome) = B0 + B1 exposure 1 + B2 exposure 2 + B3 exposure 1 * exposure 2

The test for statistical significance for B3 (parameter of interaction term) in logistic regression is the Wald Chi-square for the parameter.

Meaning of the interaction term

Predicted OR (If there is an interaction)
Log odds = B1 + B2 + B3 (additive scale)
OR = e^B1*e^B2*e^B3 (multiplicative scale)

Expected OR (without any interaction)
Log odds = B1 + B2 (additive scale)
OR = e^B1*e^B2 (multiplicative scale)

So e^B3 = Predicted OR/ Expected OR: Exponentiating the coefficient for the interaction term yields the ratio of the predicted joint effect to the expected joint effect. When the coefficient = 0 (no interaction) that ratio is 1 because e(0) = 1 so predicted joint effect = expected joint effect.

In other words,
Log predicted OR = B1+ B2 +B3
Log expected OR = B1 + B2

So B3 = log (predicted OR) - log expected OR.

Interpretation of the output for interaction in logistic regression:

- The wald chi square for the interaction term (exposure1*exposure2) is a test of the null hypothesis (there is no interaction between exposure 1 and exposure 2, so B3 is 0). It is like the breslow-day test.

- The coefficient for the interaction term is B3. E^b3 is the ratio of the predicted over the expected OR. If the ratio is < 1, then this is an antagonistic interaction (observed < expected) and if the ratio > 1 then this is a synergistic interaction (observed > expected)

- B1 and B2 (the coefficients for the main effect in the model) in the presence of an interaction term represent the effect of that variable when the other variable takes the value of zero (effect of exposure 1 among the reference level of exposure 2 and vice versa)

- You can find the predicted OR for the joint effect using Predicted = e^B1*e^B2*e^B3

- To compute the confidence interval for the predicted OR of the interaction term, make a new indicator variable and run a regular proc logistic (model outcome = strata 1 strata 2 … strata i)

Lecture 13: Case Control Studies

We will present case control studies, where the binary dependent variable is now an indicator of case-control status rather than the prevalence or absence of disease. The key difference is that we never compute the prevalence of a disease (or a prevalence OR) from a case control study: This is because the investigator is the one to decide the number of cases & Control, therefore the 'prevalence" is pre-determined. The OR in case control studies will be an eOR.

Frequency matching: Case-control design where we match a group of cases to a group of control, to make sure we have enough samples. Case control studies are great for rare diseases.

Significance Test for a 2x2 table in a case control study:

Statistical Hypothesis testing for a 2x2 table

Step 1: Formulate the null hypothesis
Ho: No association (OR=1, RR=1,RD=0)

- Calculate the probability using a chi square test
- Use the P-value for the test to accept or reject the null

Step 2: Find the test statistics

Pearson's Chi-square: Easy to use, but doesn't work with thin data.
X^2 = [observed – expected]^2/expected
X^2 = [(ad-bc)^2)*T]/[n1n0m1m0]
Degrees of freedom = (row -1)(column – 1)

- Do not use if there is less than 5 observations per cell
- For a large sample, Pearson's Chi-square and Fisher's exact test will yield the same p-value

Continuity-corrected chi-square: Chi-square alternative used for small sample sizes

Fisher's exact test : Calculated as a one tailed, but you can't just multiply the result by 2 for the two tailed result since the upper and lower tail values are different; It is always applicable, even with small cell counts

Logit Chi-square (Woolf): X^2 used on the log scale because OR is not normaly distributed (i.e, 2 and ½ are at different distance for the null which is 1) but Ln(OR) is normally distributed (Ln(1) = 0 which is the null on the log scale)

- Z = X-Xo/SE
- OR = e^B therefore, B = Ln(OR)
- Z = ln(OR) – Ln (1) / SE[ln(OR)], and **(Ln(1) = 0)**
- X^2 ~ Z^2 = [ln(OR)]^2/ SE[ln(OR)]^2
- X^2 = [ln(OR)]^2/ Var [ln(OR)] and **(SE^2 = Variance)**

- $\text{Var}[\ln(OR)] = 1/a + 1/b + 1/c + 1/d$

Miettinen Confidence Intervals: Applicable to all Chi-squares above solving for SE.

- $X^2 = [\ln(OR)]^2 / \text{Var}[\ln(OR)]$
- $X^2 = [\ln(OR)]^2 / SE[\ln(OR)]^2$
- $SE[\ln(OR)] = \ln(OR)/\sqrt{X^2}$

95% Ci for the ln(OR): $\ln(OR) \pm 1.96 * SE[\ln(OR)] = \ln(OR) \pm 1.96 * \ln(OR)/\sqrt{X^2}$

95% Ci for the OR: $e^{[\ln(OR) \pm 1.96 * \ln(OR)/\sqrt{X^2}]}$

Particular application: Logit Method (Woolf)
$\text{Var}[\ln(OR)] = 1/a + 1/b + 1/c + 1/d$
So $SE[\ln(OR)] = \sqrt{[1/a + 1/b + 1/c + 1/d]}$

95% CI for the ln(OR): $\ln(OR) \pm 1.96 * SE[\ln(OR)] = \ln(OR) \pm 1.96 * \sqrt{[1/a + 1/b + 1/c + 1/d]}$

95% CI for the OR: $e^{(\ln(OR) \pm 1.96 * SE[\ln(OR)])}$

SAS code for 2 *2 tables

```
Proc freq;
Tables x*y /chisq chm1;
/*CMH is the Cochran mantel haentzel for adjusted analysis*/
```

SAS output

- You are interested in the percents for the X variables (so if X is in row, you want row percents)
- Under relative risks, the case control CI is the logit/woolf CI (this is why the parameter is not in the middle of the confidence interval

SAS Estimates of the Relative Risk

- Case-control (Odds ratio): This is the disease odds ratio (cohort) or exposure odds ratio (case control) or prevalence odds ratio (cross-sectional).
- Cohort 1 (Column 1 risk): this is the disease risk ratio (a/a+b)
- 95% confidence bounds (asymptotic): These are the logit based confidence intervals (Woolf)
- 95% confidence bounds (exact): on the regular scale, the log transformed parameter estimate may not be in the exact middle

SAS Estimates of the Common Relative Risk: This is used for pooling estimates of a series of 2x2 tables (it will be the mantel-haenszel estimates)

SAS code for 2*n tables

```
Proc freq order = formatted;
Tables X*Y / chisq cmh1 nocol nopercent;
Format X quartile.;
```

Degrees of freedom = (row-1)(column-1)
When we have more than a 2*2 table, we cannot use the Pearson chi-square

Chi-square test of the null hypothesis of a 2*n table with (row-1)(column-1) degrees of freedom: The prevalence of disease (Y) is the same across all levels of exposures (X). Aka, there is no association between exposure and disease

The CMH (Cochran Mantel Haenszel or Mantel Haenszel Pooled Chi-square statistic) **is the 1-degree of freedom chi-square is mantel's test for trend:** it is a test for the following null hypothesis: there is no linear component to the trend across the X quartile. You perform the test by assigning an X value on an ordinal scale to each parameter. However, the test for linear trend is not a test for monotonic dose-response, but a test for monotonic trend (in the same direction)

Tabular Methods for assessment and control of confounders: For confounding, we use the etiologic model and assume that 1 exposure -> 1 disease

1. **Stratify the data on the confounder (create a 2x2 table of E and D for each stratum of the confounder)**
2. **Compute a measure of association between E and D for each stratum of exposure**
3. **Decide if it is acceptable to ignore any differences between the stratum specific measures of association (test for interaction or effect modification)**
    a. **Using the Breslow-day test:**
        - $X^2 = \Sigma (\text{observed} - \text{expected})^2 / \Sigma$ variance for each strata
        - Parameters are k-1 df, and p-value as usual
        - **If the p-value is statistically significant, then the stratum specific OR differ significantly and we should not report a summary OR.**

    b. **Logit test (Woolf)**
        - $Wi = 1 / [1/a + 1/b + 1/c + 1/d] = 1/\text{var}[\ln(OR)]$
        - $X^2 = \Sigma [\ln(ORi)^2 - \ln(ORm\text{-}h)^2] / \text{var}[\ln(ORi)] = \Sigma [\ln(ORi)^2 - \ln(ORm\text{-}h)^2] * Wi$
        - Parameters are k-1 df, p-value as usual

4. **If strata 1 ~ strata 2 then compute a pooled estimate (e.g. adjusted OR)**
    a. **Method 1: Using mantel-haenszel summary estimators**

		Disease		
		Yes	No	
Exposure	Yes	a	b	m1
	No	c	d	m0
		n1	n0	Total

Let RRi be the stratum specific measure of association and let Estimator m-h be the pooled measure of association

Estimator = Σ(Wi*RRi)/ Σ (Wi)

- For Cumulative Incidence, Prevalence or Risk Ratio: RRmh = Σ (Wi*RRi)/ Σ (Wi) = Σ (c*m1/T)*[(a/m1)/(c/m0)]/ Σ (c*m1/T) = Σ (a*mo/T)/ Σ (c*m1/T)
- For Odds Ratio: ORmh = Σ (Wi*RRi)/ Σ (Wi) = Σ (b*c/T)*[(a/b)/(c/d)]/ Σ (b*c/T) = Σ [(a*d)/T]/ Σ [(b*c)/T]

Assumptions: No residual confounding within strata because the strata is homogenized (the effect of exposure on the disease is the same within all strata because we assume no interaction between E and the confounders)

b. **using the Logit-based method**

Ln(OR) = Σ (Wi*ln(OR))/ Σ (Wi)
OR = e^ [Σ Wi*ln(OR))/ Σ (Wi)]
Wi = 1 / [1/a + 1/b + 1/c + 1/d] = 1/var [ln(OR)]

Assumptions: same as the MH estimator, but not good if the data is thin or any cell has 0, therefore the Mantel-Haenszel method is the method of choice!

5. **Compare the pooled estimate to the crude estimate**
    a. Use the 10% rule of thumb: |(lnORadj –lnORcrude)|/(lnORcrude) or (crude – adjusted)/(crude) to find out the sign of confounding
    b. If the adjusted differs from the crude by more than 10%, present the adjusted estimate
    c. Be careful about using this is the crude ratio is close to 1 (the null) or if you have a small sample size.

Unconditional Logistic Regression in Case Control Studies

eOR = (a/c)/(b/d) = Caseness OR = ad/bc

When we fit logistic regression models to case control studies, we are not modeling the log odds of disease, but the log odds of being a case versus being a control (aka the caseness odds). The log odds of the intercept is scientifically meaningless, because the prevalence (aka e^B0) is fixed by the investigator picking the case and controls. However, B1 is still useful to determine if the exposure is related to case/control status (eOR). In this case since the disease is fixed, we are interested in column percents.

The really easy way to remember that is that Pr (Y=1) = [odds Y=1]/[1+odds Y =1]

- **Probability = odds / 1+ odds**
- **Odds (of event) = probability (of event) / probability (of no event) = probability (of event)/1 – probability (of event)**

Logistic regression

- The model yields odd ratios as results
- It allows for sequential or simultaneous assessment of multiple risk factors, confounders and interaction

SAS Code

```
Proc Format;
Value Outcome 1 = ‘Yes’ 0 = ‘No”;

Proc logistic simple; /*option to get simple statistics for explanatory variables*/
Model outcome = exposure / RL;
Format outcome.;
Run;
```

Parameter estimates (B1) for X1, X2 and X3 represent the slope: This is the difference in log odds of D for the Quartile Xi compared to the reference level.

Odds Ratio are the e^B1 and estimate the odds ratio of D in Xi when compared to the reference level

*** Test for linearity: There is not a 1 degree test for trend (CMH) in proc logistic, but comparing the difference between B1 of each quartile is a qualitative estimate of the linearity assumption: This is because linarity implies that a 1 unit distance between 2 variables should be the same regardless of the position on the line, so if the values are the same then you can assume that the linearity clause holds.

95% CI = e^[B1 +/- 1.96*Se(B1)]

Wald Chi-square (Chi square in Logistic regression):

The null hypothesis is Ho: B1 =0 or Ho: OR =1
Z = B1/SE(B1)
X^2 ~ [B1/SE(B1)]^2 df = 1
Wald Chi Square: (B/se)^2

This is used as our test statistic for null hypothesis when we are looking at a single coefficient (Ho: B1 = 0)

The Log-likelihood is used for hierarchical models. This occurs when one model as all of the variables of the other and more, plus both models use the same number of observations (on the same sample)

For example
Reduced model = Model 1: B0 + B1X
Full Model = Model 2: B0 + B1X + B2Z

Model 2 is like an "extension" of Model 1. And we examine the improvement of fit by comparing them. If the full model is no better than the reduced model for fit, then the likelihood (-2log) of the full model should be close to that of the reduced model. This is testing the following null hypothesis Ho: there is no improvement of fit by adding new variables to the reduced model <-> all of the additional coefficients estimated by the full model = 0.

This Ho is tested with the likelihood chi-square test
X^2 = -2Log reduced –(-2log full) = 2logfull = 2 log reduced with df = df full – df reduced.

/* if you only have 1 exposure and you look at the improvement of fit (-2LL) the likelihood chisquare will tell you if adding the particular variable was statistically significant*/

How to calculate the OR in 2 different categories other than the reference level

Original	(B1) X1	(B2) X2	(B3) X3
Person A (Upper Category, here X3)	0	0	1
Person B (Lower Category, Here X1)	1	0	0
Difference	-1	0	1
B*Delta	B1*(-1)	B2*(0)	B3*(1)

Log Odds (Y) = B1*(-1)+ B2*(0)+ B3*(1)
OR (X3 vs X1) = exp (B3-B1) = exp (B3) / exp (B1) = OR 3/ OR 1

How to test if two level of a dummy coded (or indicator) variable are different from one another:

Method 1: test statistic by Hand

The OR to compare B1 vs B2 (Ho: OR (B1 vs B2) =1 or Ho: B1 – B2 = 0) is

OR (B1 vs B2) = Exp (B1 – B2) = exp (B1) / exp (B2)

Z = B1 – B2 / Se (B1-B2)

X^2 = (B1 – B2)^2 / Var (B1-B2)

Since Se (B1 – B2) =/= Se(B1) – Se(B2)

Var (B1-B2) = Var (B1) + Var (B2) – 2* Covar (B1,B2)

*** use the / COVB option in the model statement to get the covariance

- Use Se (B1) ^ 2 = Var (B1)
- Use Se (B2) ^ 2 = Var (B2)
- Get the Cov (B1, B2) from the /COVB option
- Find Var (B1 – B2) then Se (B1 – B2)
- Find Z then X ^2 and p-value

Exploring the Dose Response Pattern

When we are working with continuous data, the linear relationship assumption means that the shape of the association between E and D is best described by a line (at each point of the line, a 1 unit change in E will have the same effect on D and the slope is constant). However, the shape can take many forms, including concave, convex, curvilinear, cubic, etc…. The test for trend is testing for the linear component of the relation to make sure that this assumption is correct in tabular methods (proc freq). In proc logistic, we had to use indicator variables to deal with that otherwise sas would consider our ordinal strata (strata 1, 2, 3 , 4) as linear due to the numbers, and try to fit a line through it. We used the test statement to check for linearity between strata (a strata-1 df test). Therefore before fitting data as continuous in proc logistic, you should test the linearity assumption to make sure that the model is suited appropriately. This is done by fitting a quadratic term. Let's look at this example using age as exposure and cancer as outcome.

```
Proc logistic;
Model cancer = age age^2 alc smoke;
Run;
```

The term age^2 can be conceptualized as the interaction term age*age, so we are testinh that the predicted effect of age on cancer will vary across the age distribution (linearity assumption) and we follow all the regular interpretations of interaction terms. If it is significant, then the linearity assumption does not hold and the variable should not be modeled as a continuous variable in proc logistic.

Choosing the "best' model in etiologic epidemiology

In etiologic (causal) analysis, the best model is the one that can provide the following:

- A model that will provide an unconfounded estimate of the exposure-disease association
- A model that will provide a correct representation of the exposure-disease association

This is evaluated using the following steps and epidemiologic criteria:

- Substantive assessment of effect modification: using the breslow-day test or test of interaction

- Assessment of confounding: qualitative assessment comparing the crude and adjusted estimate using the 10% rule

- Improvement of fit: Comparing the reduced and full models using the LR chi Square

- Precision: This is of secondary importance because estimating the precision (aka looking at the SE for a smaller Ci) is only relevant if the estimate is unbiased (aka accurate)

- Parsimony: The more parsimonious the model, the better (aka models with less variables are simpler and better)

- Choosing the model for prediction analysis: The goal of the process is to identify a parsimonious model for the data

- Simultaneous Regression: All of the variables of interest are simultaneously entered in a single model statement

- Hierarchical Regression: This can be done by using a series of simultaneous regressions to determine whether adding each variable to the model improves the fit of the model.

- Stepwise Regression: Use subject of matter assumptions to include variables in a model, either going forward or backwards. This can be done is sas using the SLE option (significance level to leave or the p-value required to reject the variable in the model) and SLS option (significance level to stay or the p-value required to keep the variable in the model)

Class 14: Logistic Regression Goodness of fit

The Improvement of fit refers to how the inclusion of predictors or independent variables to a model improves the fit of the model to the data. The IOF is tested using the LR chi-square statistic

Goodness of fit does not focus on adding variables, but on how well a predefined model fits the data compared to all possible interactions between the variables in the model. There are 3 common tests

```
PROC LOGISTIC data = class4mkI;
class smoking (param=ref ref='0') male (param=ref ref='0');
MODEL dentist (EVENT='0') = smoking male/RL aggregate scale = none lackfit;
title 'Assesing Goodness of fit';
RUN;
```

Hosmer-Lemeshow

Partition for the Hosmer and Lemeshow Test

Group	Total	dentist = 0		dentist = 1	
		Observed	Expected	Observed	Expected
1	376	50	44.12	326	331.88
2	1172	169	160.13	1003	1011.87
3	227	35	40.88	192	186.12
4	974	193	201.87	781	772.13
5	477	99	113.74	378	363.26
6	473	176	161.26	297	311.74

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
7.5321	4	0.1103

This is computed by specifying lackfit as an option in the SAS statement: the null hypothesis is that the model fits the data. The test is a chi-square (observed vs expected) computed in the following way:

- Predicted probability of the outcome is computed for each observation
- Observation are sorted by their value of predicted probability
- Observations are grouped in 10 intervals of the predicted probability
- The predicted probabilities are added to create the predicted (or expected) values at each interval
- Expected values are compared to the observed values and the degrees of freedom are the number of intervals minus 2.

Deviance

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	**Value**	**DF**	**Value/DF**	**Pr > ChiSq**
Deviance	7.5912	2	3.7956	0.0225
Pearson	7.5321	2	3.7661	0.0231

This is computed by aggregate scale = none in the model statement. This is similar to the -2log-likelihood by comparing our model to the ‘saturated’ model (model with all of the possible interactions between the variables, even if they are not statistically significant) by comparing the 2 LR between our model and the saturated models. The null hypothesis is that our model fits the data as well as the saturated model. If the deviance is large, then our model’s fit is worse than the saturated models fit. The larger the deviance (difference in LR chi squares) the worse the model. In other words, the smaller the p-value, (usually < 0.2), the worse the model.

C-test

Responses			
Percent Concordant	50.9	**Somers' D**	0.220
Percent Discordant	28.9	**Gamma**	0.276
Percent Tied	20.1	**Tau-a**	0.069
Pairs	2149394	**c**	0.610

Sensitivity (true positive)
False positive = probability of incorrectly predicting an event = 1 – specificity = 1- TN

This is comparing the area under the curve by graphing the sensitivity to the false positives: at each cut point in the x variable, calculate the sensitivity and false positive rate, plot them and create the ROC curve. This can also be done in SAS:

```
ODS Graphics on;
Proc logistic proc = roc;
ODS Graphics off;
```

Lecture 16 Multinomial Regression

So far in this course, we have considered both continuous and dichotomous outcomes. Ordinal outcomes are those with multiple categories that can be ordered. Nominal Variables are variables that have multiple categories but they cannot be ordered. It is tempting to analyze ordinal data using indictor variables.

If you categorize and treat the variable as dichotomous, we can use logistic regression or multiple logistic regression models to compare the different categories. However logistic regression will discard data about the ordinality of the data, and dichotomizing the data is often arbitrary, leading to homogenization of data that is assumed to be equal in the linear equation.

If you treat the data as continuous and use linear regression, you assume homogeneity in the variance for the outcome variable, but this is not accurate since we have different categories.

It is better to use models that take ordering into account: They are easier to interpret, and Hypothesis testing is more powerful since we do not ignore ordering.

Proportional Odds Model

Whenever proc logistic encounters more than two categories on the dependent variable (outcome) it estimates a cumulative logit model:

Odds (Y=1) = (Pr Y=1)/(1- Pr Y=1) = Bo + B1*X
Pr (Yj = 1) = e^ Boj + B1*X/ [1+ e^Boj + B1*X]

And j represents the levels of the outcome.

- In logistic regression, P (Y=1) + P(Y=0) = 1
- In Multinomial regression P (Y =1) + P (Y=2) + … + P (Y=j) =1
- The cutpoint specific estimates are not statistically independent

The key assumption is that the log odds ratio for proportional odds/cumulative logit are identical across each cutpoint in our outcome, and each cutpoint has its own associated intercept.

Ho: Test whether were the cutpoint is made matters (OR1=OR2=OR3) for proportional odds. For example, the number of outcome categories is more than 2, or the number of exposure categories is the product of each number of strata in variables

SAS Code

```
Proc logistic descending;
Model Y = X;
Run;
```

- The logistic model will run a cumulative Logit when the outcome has multiple levels
- The Score test for the proportional odds assumption is testing whether OR1=OR2=OR3. If we accept the null, we can use the cumulative odds ratio. Therefore this is testing the appropriateness of the model. The number of DF is p variables * (k-2) outcome strata.
- The -2LL is testing the model fit (intercept vs intercept + covariates)
- The Wald Chi Square is testing if B is statistically significant.
- E^B will give us our Odds Ratio
- Each intercept represent the log odds of disease among the unexposed for each outcome category. This is of limited use in case control studies since we are setting the rate of disease by nature of our study design

Model	Outcome Coding	Underlying Distribution	Measure of Association
Linear	Continuous	Gaussian	Mean Difference
Proportional Odds	Ordinal	Multinomial	Cumulative OR
Logistic	Binary	Binomial	OR

Polytomous Logistic Regression

If the proportional odds assumption is badly violated, you have to run a polytonomous model to model the outcome with multiple levels

SAS Code

```
Proc Logistic Descending
Model Y = X/ link = glogit;
Run;
```

There will be no score test for proportional odds, since the polytomous model assumes that the proportional odds assumption is violated. There will be separate parameter estimates for both the intercept and each level of the outcome being considered.

Class 17: Matched Case Control Studies (Classical Methods and Conditional Logistic Regression)

Matching is generally only done in case control studies, and then only in studies with a small sample size

Matched Analysis: Individual Matching

		Control		
		Exposed	Unexposed	
Case	Exposed	A	B	A+B
	Unexposed	C	D	C+D
		A+C	B+D	N

The cells are different from the cell entries in the unmatched analysis
The number within each cell is the number of cells, not the number of individuals
A and D are the concordant Cells, and do not contribute to the OR
B and C are the discordant Cells, and contribute to the OR

Why is the OR B/C?

In the case of individual matching (1:1 Matching) we are actually stratifying the data so that each stratum only contains 2 observations (the case and its control). The stratum will have one of the following patterns:

A Pairs

	CA	CO
Exposed	1	1
Unexposed	0	0

B Pairs

	CA	CO
Exposed	1	0
Unexposed	0	1

C Pairs

	CA	CO
Exposed	0	1
Unexposed	1	0

D Pairs

	CA	CO
Exposed	0	0
Unexposed	1	1

Then, we compute a Mantel-Haenszel summary estimator across all of the strata:

ORmh = Σ [(a*d)/T]/ Σ [(b*c)/T]

pairs (A pairs, D pairs) do not contribute to either the numerator (ai*di/Ti = 0) or the denominator (bi*ci/Ti = 0)

Each discordant pair (B type) contributes ½ to the numerator (1*1/2) and 0 to the denominator (0*0/2)

Each discordant pair (C type) contributes 0 to the numerator (0*0/2) and ½ to the denominator (1*1/2)

Therefore ORmh = Σ [(a*d)/T]/ Σ [(b*c)/T] = ½*B / ½*C = B/C

Measure of association: ORmh = B/C
Statistical Hypothesis testing (McNemar's Test) for Ho: OR =1
McNemar's X^2 (B and C >5) = (B-C)^2/(B+C)
McNemar's Continuity Corrected X^2 (B or C <5)= (|B-C|-1)^2/ (B+C)
Both are Chi-square with 1 df.

95% CI
Method 1: 95% CI = exp [lnOR +/- 1.96*SE(lnOR)]
Var (lnOR) = (1/B + 1/C)
SE (lnOR) = Sqrt (Var(lnOR))

Method 2: 95% CI = exp [lnOR +/- 1.96*(lnOR)/sqrt X^2]

Variable Number of Controls per Case (1:C rather than 1:1)

- Measure of Association: Matched OR (Separate each matched set into its 2x2 and compute ORmh)
- Statistical Hypothesis Testing: Compute X^2 as in stratified unmatched analysis
- Confidence Interval: Compute the Test based 95% CI using X^2
- Efficiency (how many control per cases can we recruit) this is determined using 2C / (C+1). As C tends to infinity, the efficiency tends to 2. It usually plateaus around C = 5

Using SAS Proc Freq

Since the matched OR is a ORmh, we can use proc freq.
Define a variable SET that indicates the number of each matched set (to stratify the data)
Use the following model in SAS: CxExD, and use the noprint option so that sas does not print a 2x2 for every single pair

```
Proc freq;
Tables SET*EXPOSURE*DISEASE/ noprint CMH1;
```

The CMH chi square will be the equivalent of the MCNemar's Chi Square
The Breslow day test is not a useful statistic (tests for the homogeneity across strata) because we have a LOT of strata. However, the # of df is k-1 the number of pair in our dataset

Using SAS Proc Logistic (conditional logistic Regression)

The user must tell sas which subjects are grouped together as matched sets

```
Proc logistic;
Strata matchnum;
Model case (event = "1") = exposure;
Run;
```

Review of the output

- You can tell that it is a conditional logistic regression model because there is no intercept term
- Three test for Ho: OR = 1 (Likelihood X^2, Score X^2, Wald X^2). All tests are based on 1 DF
- 95% CI: exp [lnOR +/- 1.96*lnOR]
- you can use the exactonly option after proc logistic to run an exact conditional logistic regression

Residual Confounding
Even though matching cases to control in theory takes care of confounding, we can still have residual confounding for continuous variables, since we are homogenizing data across a range. You can include the continuous variables in the model, and use the 10% rule to make sure that there is no residual condounding.

Lecture 20: Introduction to Cohort Studies, Cumulative Incidence, Relative Risk Regression, Incidence Density

Cohort Studies: Measure exposure status at baseline and identify disease occurrence over follow-up time

Cumulative incidence: One counts events and non events occurring in the exposed and unexposed, computes event risks (cumulative incidence) and thus the risk ratio associated with the exposure (Cumulative Incidence Ratio or Relative Risk)

Incident Density Methods: one computes the event rates using person-time as the denominator and thus, the rate ratio (incident density ratio)

The simplest way to show the relationship between 2 binary variables is a 2 x 2 table

		Disease		
		Yes	No	
Exposure	Yes	a	b	m1
	No	c	d	m0
		n1	n0	Total

- For Cohort studies, the risk ratio is the prevalence ratio: [a/a+b]/[c/c+d]
- **The risk difference [a/a+b] - [c/c+d] represents the excess risk**
- For Cohort studies, the risk difference gives us the attributable risk due to exposure and is: [a/a+b] - [c/c+d]
- For Cohort Studies, the Odds ratio is the ratio of 2 odds and can be expressed in 2 ways
  - Disease odds (cohort): Disease odds among exposed / Disease odds among unexposed = [a/b]/[c/d]
  - Exposure Odds (case control): Exposure odds among diseased / Exposure odds among non-diseased = [a/c]/[b/d]
  - Disease odds = Exposure odds = (a*d)/(b*c)
  - **When the disease is rare (a is low and c is low) the risk ratio ~ odds ratio**

Manteh Haenszel Summary Estimators

RR mh = Σ [(ai*moi)/Ti] / Σ [(ci*m1i)/Ti]

95% CI: e^(ln(OR) +/- 1.96*SE[ln(OR)]) = e^(ln(OR) +/- 1.96 * √[1/a +1/b + 1/c + 1/d])

*** The Test of Homogeneity is not the same as the breslow day test

Logistic Regression for Cumulative Incidence Data

Odds = e(Bo + B1x)
Pr (Y=1) = Odds / (1+Odds)
Pr (Y=0) = Odds / (1+Odds)
RR = Pr (Y=1) / Pr (Y=0)

Interpreting The Coefficients

Bo is the background odds of the disease
To assess for confounding, fit another model for the confounding variable and compare the odds ratio.

Relative Risk Regression

Start With Linear Probability Model: Pr (Y=1) = Bo + B1X -> Put it in log scale (so that the difference between Y =1 and Y = 0 is a ratio) Log (Pr Y) = Bo + B1x.

Use Proc Gen Mod with a log link and a binomial distribution for our outcome instead of proc logistic

```
Proc genmod;
Model Y = X / ling = log dist = bin;
Run;
```

The difference between proc logistic and proc genmod log linked is that proc logistic uses the link logit: (odds = Pr / (1-Pr) or (Pr = Odds / 1 + Odds).

The RR estimates from proc genmod are more conservative than the ORs from proc logistic.

Lecture 21: Cohort Studies and Poisson Regression

Unconditional Logistic Regression can approximate cumulative incidence, and relative risk can directly estimate cumulative incidence, but neither allow variable time-to-event data to be entered

	Poisson Regressi0n	Logistic Regression
Dependent Variable	Log (Rate)	log (odds)
Measure of Association	Rate Ratio	Odds Ratio
Person-Time At Risk	Considered	Ignored

Rate = r events / Person Time

Log (rate) = log (r events / Person Time) = B0 + B1x
Log (r events) – log (person time) = B0 + B1x
Log (r events) = B0 + B1x + log (Person time)

The Log (Person time) term is accounted for as an offset term in the proc genmod code

```
Proc genmod;
Model y = x / dist = poisson link = log offset = person time type3;
Run;
```

The computation of the rate ratio never includes the intercept term
Rate = exp (B0 + B1x)

The type3 option in the model gives you a Likelihood ratio statistis for type 3 analysis for the covariates in the model.

The deviance statistic for poisson regression is not useful to evaluate goodness of fit

The model's Log likelihood is reported, so you have to manually compute the -2 Log Likelihood to compare models

Lecture 22: Poisson regression, Continued

```
/*Calculating person-time for your exposure of interest*/

/*Method 1: Counting and using the retain function to go on longitudinal data starting at
the first observation until the last observation*/

data framinghammkI;
set framingham;
if sex=2;  /* male=0 */

retain totalpy 0; /*start by zeroing your count*/
totalpy=totalpy + miyears;

proc means mean n min max; /*Max is the total person time*/
var totalpy;
title1 'data for females';
run;

proc freq; /*number of event*/
tables hospmi;
title1 ‘events divided by person-time is incidence rate’;
run;

/*rate ratio = number of event / total person time = freq / max*/

/*Method 2: Using the sum option in PROC MEANS*/

proc sort data=temp; by sex;

proc means data=temp sum;
var miyears;
by sex;
run;

/*If you want to save these counts for use in other parts of your coding, the data can be
outputted to a new temporary dataset with the OUTPUT statement*/

/*Question 1*/

/*sum up the person time by gender*/

proc sort data=framinghammkI; by male;

proc means data=framinghammkI sum;
var miyears;
```

```
by male;
title 'person time';
run;

proc freq data = framinghammkI; /*number of event*/
tables hospmi;
by male;
title1 'event numbers';
run;

/*Question 2*/

proc genmod data = framinghammkI;
model hospmi=male/dist=poisson link=log offset=logmi type3;
title 'poisson regression with log time offset';
run;

proc genmod data = framinghammkI;
model hospmi=male/dist=poisson link=log offset=miyears type3;
title 'poisson regression with regular time offset';
run;

/*should use log of person time because poisson regression is based on log (events) = Bo
+ B1x + Log (Person-time)*/

/*Question 3*/

proc genmod data = framinghammkI;
model hospmi=male age/dist=poisson link=log offset=logmi type3;
title 'poisson regression with log time offset adjusted for age';
run;

/*Question 4*/

proc genmod data = framinghammkI;
model hospmi=male age male*age/dist=poisson link=log offset=logmi type3;
title 'poisson regression with log time offset testing for interaction';
run;

/*Question 5*/

proc genmod data = framinghammkI;
model hospmi=male age male*age age*age male*age*age/dist=poisson link=log
offset=logmi type3;
title 'poisson regression with log time offset testing for interaction of quadratic age';
run;
```

```
/*have to keep male*age in model because the interaction was statistically significant in
question 4*/

/*Question 6*/

proc genmod data = framinghammkI;
model hospmi=male age age*age/dist=poisson link=log offset=logmi type3;
title 'poisson regression with log time offset testing for quadratic age main effect';
run;

/*PROC GENMOD can also generate logistic regression and relative risk (log-binomial)
regression models by changing the link and distribution options. Recall that if you have
cohort data, logistic regression estimates a cumulative incidence odds ratio and relative
risk regression estimates relative risk or risk ratio*/

/*Reminder of these options*/

/*Relative risk regression: link=log dist=bin*/
/*Logistic regression: link=logit dist=bin*/

/*Question 7*/

proc genmod data = framinghammkI descending;
model hospmi=male/link=log dist=bin type3;
title 'Relative risk regression';
run;

proc genmod data = framinghammkI descending;
model hospmi=male/link=logit dist=bin type3;
title 'logistic regression';
run;

/*Question 8*/

proc genmod data = framinghammkI descending;
model hospmi=male age/link=log dist=bin type3;
title 'Relative risk regression with age as a confounder';
run;

proc genmod data = framinghammkI descending;
model hospmi=male age/link=logit dist=bin type3;
title 'logistic regression with age as a confounder';
run;
```

Lecture 23: Survival Analysis

There are four methods to deals with incidence density data:

- Tabular Methods
- Life Tables
- Poisson Regression
- Cox Regression

Survival analysis is a collective term for statistical methods that are used to study binary outcomes such as dead and alive while taking into account the time when the event of interest occurs. We chose not to use logistic regression for this type of analysis for two main reasons: logistic regression doesn't take into account varying person-time at risk, and the measure of association from logistic regression is an odds ration, not a relative risk.

There are 3 types of time at risk to consider:

- Observation start – Event observed
- Observation start – lost to follow up
- Observation start – end of the study

Survival function

T is defined as the time to event for a given person

For any time point t, F(t) = Pr (T<t). F is probability between 0 and 1.

In survival analysis, it is more common to work with the survivor function:

S(t) = 1 – F(t). This is the probability of not experiencing the event, and is a stepwise decreasing function

Kaplan Meier (KM) Method

	T0	2 deaths	T1	1 death	T2	1 Death and 1 Censored	T3
No At Risk (At beginning of the interval)	25		23		22		20
Pr (death) at that time point	0/25		2/25		1/23		1/21*
Pr (Survival) in the interval	1		1- (2/25)		1-(1/23)		1-(1/21)
S(t) cumulative	1		1-(2/25)		[1-(2/25)]*[1-(2/23)]		[1-(2/25)]*[1-(2/23)]*[1-(1/21)]

Divide the number of events by the number of people at risk at the beginning of the previous interval (aka people at the beginning of interval)

The cumulative survival rate is the product of the survival probability at all intervals

If there are censored in the interval, then those need to be removed from the number at risk

SAS Code

```
Proc lifetest method = km;
Time dur*status(0);
Strata k;
```

The first variable dur is the time at risk variable.
The second variable event is for the event status. The number in parenthesis is the vale of 'no-event'

*** the number failed column is the cumulative number of event, so you need to look at the difference between each interval

To test for homogeneity of survival curves over strata, you need to look at the log-rank and Wilcoxon chi squares . You have to use the log rank test because the Wilcoxon gives more weight to the observations at the beginning of study.

If you use

```
Proc lifetest method = life plots=s;
Time dur*status(0);
Strata k;
```

Then you will looking at interval specific times instead of cumulative events.

Lecture 25: Cox Regression I

Rate is defined as #events/person-time at risk. The denominator can be conceptualized as a single person who is observed (at risk) for some period of time where time is measured in some specified unit.

Now assume that we are interested in the instantaneous rate of the disease at a time point t: it is often referred to as the hazard rate, and its related measure of association is called the hazard rate.

Cox regression is modeling the log(rate) but does not assume a constant rate over time, and allows for a continuously changing hazard function

Cox regression can be considered a semi-parametric method (parametric methods assume that there is some distribution from which the calculations are derived)

h(t) = [reference hazard]* e ^[B1 * X1]

The reference hazard is a function of time that could change continuously (ie over the follow up time)

h(t) = ho(t) * e ^[B1 * X1]

log [h(t)] = log [ho(t)] + [B1 * X1]
log [h(t) Person a/h(t) Person b] = [B1 * X1]

B1 = Predicted Difference in log(hazard) per one unit increment in X

Since log [h(t) Person a/h(t) Person b] = [B1 * X1]
Is equivalent to =log [h(t) Person a] – log [h(t) Person b] = [B1 * X1] which is a constant, the two time functions should be parallel over time

SAS Code

```
Proc PHREG;
Model follow-uptime*outcome(censored value) = exposure;
Run;
```

Review of the output
Dependent Variable : SAS will label the follow-up time as the dependent variable but that is not correct (this is actually the time at risk). Remember that the dependent variable in cox regression is log(rate of disease) so just like in logistic regression, the real dependent variable is incident disease (aka the censoring variable)

Ties handling:means that two or more observations have the same event time (follow up time) the default is ties = breslow. The ties = efron option is often recommended

Testing Global Null Hypothesis: similar to proc logistic, score test here is the same as the log rank test

The maximum likelihood estimates does not have an intercept term (characteristic of Cox regression)

1 – the percent censored = failure rate

Testing the proportional hazards assumption

We can test the assumption by modeling an interaction between exposure variable and time at risk

$\log [h(t)] = \log [ho(t)] + [B1 * X1] + B2*t*X1$

- if B2 is not statistically significant, we do not have enough evidence against the proportional hazards ratio
- if B2 is statistically significant we have statistical evidence that the proportional hazards ratio assumption does not hold
- if B2 is positive, the hazard ratio increases over time
- if B2 is negative, the hazard ratio decrease over time

SAS code

```
Proc PHREG;
Model follow-uptime*outcome(censored value) = exposure;
Phazard = exposure*[log(follow-uptime)]
Run;
```

Interpretation of phazard is similar to that of test for homogeneity of the hazard ratio over time (B2)

Lecture 26: Cox Regression Part II

```
libname epi3 "C:\Users\SPH-User\Desktop\SAS";

data framingham;
set epi3.newframingham;

proc contents data=framingham;
title1 'contents of class6 dataset, with labels';
run;

proc print data = framingham (obs=50);
title 'looking at the class data';
run;

/*data cleaning, remove all prevalent mi since we are doing a cohort study, rename sex
and changing */

data framinghammkI;
set framingham;
if sex = 1 then male = 1;
if sex = 2 then male = 0;
years = time/365.25;
apyears = timeap/365.25;
chdyears = timechd/365.25;
cvdyears = timecvd/365.25;
dthyears = timedth/365.25;
hypyears = timehyp/365.25;
miyears = timemi/365.25;
mifcyears = timemifc/365.25;
strkyears = timestrk/365.25;
logstrk = log (strkyears);
if prevstrk = 1 then delete;
if prevstrk = . then delete;
if stroke = . then delete;
if male = . then delete;
if timestrk = . then delete;
if age = . then delete;
if timestrk le 0 then delete;
run;

proc contents data = framinghammkI;
run;

proc print data = framinghammkI(obs=50);
run;
```

```
/*Question 2*/
proc means mean n min max;
var age strkyears;
run;

proc ttest; /*continuous variable*/
class stroke;
var age strkyears;
title 'comparing the means age and strkyears between stroke = 1 and stroke = 0';
run;

proc sort; by descending male descending stroke; /*2x2 table format*/

proc freq order=data;
tables male*stroke/chisq; /*categorical variable*/
title 'testing if proportions are different from one another';
run;

/*Question 3*/
data framinghammkI;
set framinghammkI;
if age ge 50 then agegrp = 1;
else agegrp = 0;
run;

PROC LIFETEST GRAPHICS NOTABLE PLOTS=S (NOCENSOR ATRISK=0 to 24
by 6);
TIME STRkYears*STROKE(0);
STRATA AGEGRP;
run;

/*to plot the curve for failure instead of the survival plot, use the following code*/
PROC LIFETEST GRAPHICS NOTABLE PLOTS=S (NOCENSOR ATRISK=0 to 24
by 6 failure);
TIME STRkYears*STROKE(0);
STRATA AGEGRP;
run;

/*Notes:
PLOTS=S: asks for the survival plot
NOTABLE: suppresses the display of survival function estimates for each subject. Only
the number of censored and event times, plots, and test results are displayed.
NOCENSOR: removes censoring symbols or tick marks from the plot
ATRISK=0 to 24 by 6: requests that the number at risk be displayed at the time points
specified. In this example, the number still at risk will be displayed at 6-year intervals,
from year 0 to year 24, which is the maximum length of follow-up.*/
```

```
/*Question 4*/
/*Cox regression code*/
PROC PHREG data = framinghammkI;
MODEL STRKYEARS*STROKE(0) = AGE / RL TIES=EFRON;
RUN;

PROC PHREG data = framinghammkI;
MODEL STRKYEARS*STROKE(0) = AGE / RL TIES=BRESLOW;
RUN;

PROC PHREG data = framinghammkI;
MODEL STRKYEARS*STROKE(0) = AGE / RL TIES=EXACT;
RUN;

/*Question 5*/
PROC PHREG data = framinghammkI;
MODEL STRKYEARS*STROKE(0) = AGE AGE*AGE/ RL TIES=EFRON; /*This
will not give you hazard ratio, you should code for age within the data step because
causing directly in the model step does not allow time to move forward, instead it creates
a constant*/
RUN;

/*Appropriate model*/
PROC PHREG data = framinghammkI;
MODEL STRKYEARS*STROKE(0) = AGE agesq/ RL TIES=EFRON; /*This will
allow you to accumulate the time variable over the follow up time*/
agesq = age*age;
RUN;

/*Question 7*/
PROC PHREG data = framinghammkI;
MODEL STRKYEARS*STROKE(0) = AGE MALE agemale/ RL TIES=EFRON;
agemale = age*male;
RUN;

/*Question 8*/
PROC PHREG data = framinghammkI;
MODEL STRKYEARS*STROKE(0) = AGE PHAZARD / RL;
PHAZARD = AGE*(LOG(STRKYEARS));
TITLE1 ‘PHAZARD IS THE TEST OF THE PROPORTIONAL HAZARDS
ASSUMPTION’;
RUN;

/*Option 2: the age will not be properly calculated outside of the model, so you should
use it within the model statement: aka option 1 is the correct 1*/
```

```
data framinghammkI;
set framinghammkI;
PHAZARDtime = log(strkyears);
run;

PROC PHREG data = framinghammkI;
PHAZARD = AGE*PHAZARDtime;
MODEL STRKYEARS*STROKE(0) = AGE PHAZARD / RL;
TITLE1 ‘PHAZARD IS THE TEST OF THE PROPORTIONAL HAZARDS
ASSUMPTION’;
RUN;

/*Question 9*/
/*Divide follow-up time at the midpoint (0-12, >12-24 years) with 0-12 years as the
reference group, then create variables reflecting the interaction of age and time interval.
As with the interaction term, above, you will need to create the indicator variables for
follow-up time within the PHREG procedure itself. Please note that this method of
dividing follow-up time into discrete windows will not work if variables are created
within the DATA step*/

PROC PHREG;
MODEL STRKYEARs*STROKE(0) = AGE PH12_24;
IF 0 LT STRKYEARS LE 12 THEN TIME1=1; ELSE TIME1=0;
IF STRKYEARS GT 12 THEN TIME2=1; ELSE TIME2=0;
PH0_12 = AGE*TIME1;
PH12_24 = AGE*TIME2;
RUN;

PROC PHREG;
MODEL STRKYEARS*STROKE(0) = AGE PH9_20 PH20_24;
IF 0 LT STRKYEARS LE 9 THEN TIME1=1; ELSE TIME1=0;
IF 9 LT STRKYEARS LE 20 THEN TIME2=1; ELSE TIME2=0;
IF STRKYEARS GT 20 THEN TIME3=1; ELSE TIME3=0;
PH0_9 = AGE*TIME1;
PH9_20 = AGE*TIME2;
PH20_24 = AGE*TIME3;
RUN;
```

Lecture 28: Correlated Data Analysis

There are typically three types of statistical models that we commonly use in epidemiologic data analyses

	Distribution	Link	Mean Function	Variance Function	Estimate
Gaussian	Normal	Identity	g(u) = u	var(u) = 1	Risk Difference
Logistic	Binary	Logit	g(u) = u/(1-u)	var(u)= u(1-u)	Odds Ratio
Poisson	Binary	Log	g(u)=log(u)	var(u)=u	Relative Risk

Correlated data looks at cases where the observations are not independent from each other

If you ignore the correlation and analyze the observations as independent, you can have incorrect inferences.

If you use a summary measure for the cluster (average/ post – pre differences) you loose power by not using all of the information

Observations can be correlated within a person (repeated mesaures) or between individuals (correlated clusters)

Analysis of correlated data is best done in long format: each row of data is a separate observation, so each individual has more than one row

Correlation structures: when you are looking at correlated data, there are canonical structure that can be used to describe how observations are correlated

Independence Matrix: Assume independence
Exchangeable /Compound Symmetry: Common variance as each point, with common covariance too
AR1: Common Variance as each observation, but covariance decays exponentially with time so that the correlation value between two observations equals the baseline correlation value raised to the power equal to the absolute difference between the responses
Toepliz: Common variance at each time with a unique covariance for each covariance (banded)
Unstructured: Each correlation is individually measured

Correlation related to covariance: Correlation = covariance (a,b)/var(a)*var(b)

Types of correlated data analyses

SAS Proc Mixed = continuous fixed & Random effect (within cluster effect)

SAS Proc Genmod = continuous or binary fixed effect (between cluster effect)

SAS Proc Glimmix = binary outcome with random effects (combine within and between cluster effect for binary/discrete outcomes)

SAS CODE

```
Proc Mixed data = dataset;
Class (cluster variable);
Model Y = X / s; use s for solutions and parameter estimates when using the class variables*
Repeated (cluster variable) / type = (correlation structure) rcorr (prints the correlation);
```

```
Proc Genmod data = dataset;
Class (cluster variable);
Model Y = X / dist = (distribution) link = (function link) type3 wald;
Repeated (cluster variable) / type = (correlation structure) corrw(prints the correlation);
```

** if you do not include the type = corr option in genmod, SAS will print the estimated correlation structure that best fits your data