Class 8 Outline - Midterm Review

1. Sample size estimation
2. Review of generalized linear models
3. Propensity score strategies
4. Survival analysis
   - Ungrouped survival data
     - Kaplan-Meier estimate of the survival function, S(t)
     - Log-rank test for comparing survivor curves
     - Cox proportional hazards regression
5. Interpretation of Cox regression coefficients
6. Review of inference from regression models: estimates, p-values, 95% CIs, nested models

1

0. Learning Objectives

- Distinguish assumptions made in estimating sample size for purposes of precision versus hypothesis sample and identify required components.
- Define generalized linear models.
- Describe the rationale for propensity score strategies to adjust for possible confounders.
- Use survival analysis to compare survival experience of groups and interpret the results in substantive terms.
- Use multiple regression models and interpret the results in substantive terms.

2

1. What distinguishes one sample size calculation from another?

- Precision: Able to **estimate** to within +/- d units or percentage points assuming a specified alpha
- Hypothesis **testing**: Able to "detect" (as statistically significant) a difference defined by the null hypothesis vs. the alternative hypothesis with specified alpha and **power**
- Each type of sample size calculation may be for one group or for the difference between two groups

3

1.1 Sample Size for Precision

- Can calculate sample size from the point of view of estimation→ concern with precision (width of resulting confidence interval)
- w= width of the confidence interval
- w/2 = ½ width = d such that $d = Z_{\alpha/2} \cdot SE$
- Set α; solve for n

4

1.2a Sample Size for Hypothesis Testing

Define:

α = Prob (rejecting $H_o | H_o$ true)

β = Prob (failing to reject $H_o | H_a$ true)

$1-\beta$ = Power of a statistical test

= Prob (rejecting $H_o | H_a$ true)

Power is calculated for a **specific** value of Δ

Factors influencing sample size: α, β, $(1-\beta)$, Δ and variance

5

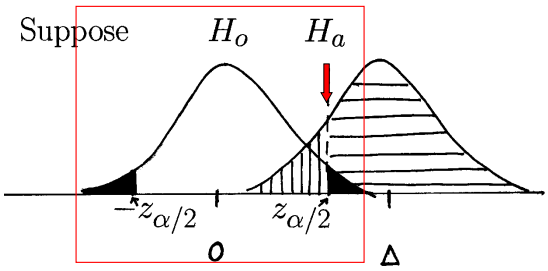
1.2b Sample Size for Hypothesis Testing

- Continuous outcome versus dichotomous outcome
- One sample versus two samples
  - Equal sample sizes per group
  - Unequal sample sizes

6

1.3a Review of Hypothesis Testing

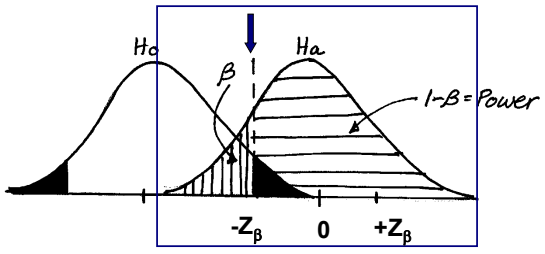
- Define Ho and Ha
- Set α
- Reject Ho when $Z_{obs} > Z_{\alpha/2}$ or $Z_{obs} < -Z_{\alpha/2}$ <u>under the assumption that Ho is true</u>



7

1.3b Review of Hypothesis Testing

- For a particular Ha, we use the critical value as the cut point for determining β (and power) <u>under the assumption that Ha is true</u>



8

1.4 Sample Size for One Sample – Hypothesis Testing

Population Value	Estimator	Sample Size
μ	$\overline{X}$	$n = \dfrac{(z_{\alpha/2} + z_{\beta})^2 \sigma^2}{\Delta^2}$
p	$\hat{p}$	$n = \left[\dfrac{z_{\alpha/2}\sqrt{p_0 q_0} + z_{\beta}\sqrt{p_a q_a}}{\Delta} \right]^2$

9

1.5 Sample Size for Two Samples (Equal Sample Sizes)

- Assuming equal sample sizes $n_1 = n_2 = n$

Population Value	Estimator	Sample Size
$\mu_1 - \mu_2$	$\overline{X}_1 - \overline{X}_2$	$n_1 = n_2 = \dfrac{(z_{\alpha/2} + z_{\beta})^2 (\sigma_1^2 + \sigma_2^2)}{\Delta^2}$
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$n_1 = n_2 = \dfrac{\left(z_{\alpha/2}\sqrt{2\overline{pq}} + z_{\beta}\sqrt{p_1 q_1 + p_2 q_2}\right)^2}{\Delta^2}$

10

1.6 Sample Size for Two Samples (Unequal Sample Sizes)

- Assuming unequal sample sizes $n_2=\lambda n_1$

Population Value	Estimator	Sample Size
μ_1- μ_2	$\overline{X}_1 - \overline{X}_2$	$n_1 = \frac{(z_{\alpha/2} + z_\beta)^2(\sigma_1^2 + \sigma_2^2/\lambda)}{\Delta^2}$ $n_2 = \lambda n_1$
p_1-p_2	$\hat{p}_1 - \hat{p}_2$	$n_1 = \frac{\left(z_{\alpha/2}\sqrt{\overline{pq}(\lambda+1)\lambda} + z_\beta\sqrt{p_1q_1 + p_2q_2/\lambda}\right)^2}{\Delta^2}$ $n_2 = \lambda n_1$

11

1.7a Summary of Sample Size Considerations

- Sample size can be calculated for one or two groups for purposes of :
  - Precision (requires desired width of CI and specification of α)
  - Hypothesis testing (requires null hypothesis and specific alternative hypothesis [to calculate Δ] and specification of assumed α,β, variance)
- Sample sizes for two groups:
  - Equal samples sizes $n_1=n_2=n$
  - Unequal samples $n_1 \neq n_2$

12

1.7b Summary of Sample Size Considerations

- Sample size is derived from knowledge of the sampling distribution of the sample statistic of interest
- Sample size determination must go beyond calculating a single value
- Choice of sample size depends on a balance of reasonable assumptions, time, effort, and expense
- Statistical significance versus practical significance

13

2. Review of Generalized Linear Models

Generalized Linear Models (GLMs)
provide a way to express the relationship between the response (Y) and explanatory variables (X's):

$$\text{Function of expected } Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

The β's, the "regression coefficients" express the relationships.

14

2.1 Types of Generalized Linear Models

Model	Response	Distribution	Regression Coefficient Interpretation	Link Function
Linear	Continuous	Gaussian	Change in ave(Y) per unit change in X	μ
Logistic	Binary	Binomial	Log odds ratio	log (odds)
Log-linear	Times to events (counts)	Poisson	Log incidence rate ratio (log IRR)	log(incidence rate)
Proportional hazards	Times to events	Semi-parametric	Log hazard ratio (log HR)	log(hazard rate)

15

2.2a Regression Models

- Linear regression model with p covariates

  $Y_i = E(Y_i \mid X's) + \varepsilon_i$

  $= E(Y_i|X's) = \mu_i = \beta_0 + \beta_1X_1 + \beta_2X_2 + \ldots.. + \beta_pX_p$

- Logistic regression model with p covariates

  $Y_i \sim \text{Binomial}\,(E(Y_i \mid X_i) = p_i)$

  $\log \text{odds}\,(\Pr(Y_i=1)) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \ldots.. + \beta_pX_p$

- Cox model with p covariates

  $\log h(t, X) = \log h_0(t) + \beta_1X_1 + \beta_2X_2 + \ldots\ldots + \beta_pX_p$

16

2.2b Regression Models

- Log-linear (Poisson) model with p covariates

  $Y_i \sim \text{Poisson}(\mu_i)$

  $\log(\mu_i) = \log N_i + \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_p X_p$

  $\log(\lambda_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_p X_p$

  More to come!

17

3.1a Propensity Scores to Control for Potential Confounding

- To estimate the effect of a "treatment" or "risk factor" (e.g., smoking) on an outcome (e.g. major smoking caused disease) by comparing "otherwise similar" persons with and without the treatment.
- Controlling for one covariate:
  - Stratify by the covariate
  - Estimate the difference in mean outcome or log odds ratio within each covariate stratum
  - Pool the stratum-specific estimates of effects absent evidence of qualitative effect modification

18

3.1b Propensity Scores to Control for Potential Confounding

- Controlling for <u>many</u> covariates:
  - Stratify on all confounder combinations (large number of strata, hard to make tables)
  - Match each smoker to a few "similar" non-smokers; not bad, but does not use all the data
  - Stratify on a single derived variable chosen so that the distribution of all the covariates is similar for the two treatment groups within each stratum of the variable.
  - One such variable is the **propensity score**

19

3.2 Definition of Propensity Score

- Definition: $p(X) = Pr(Z=1|X)$
  - The propensity score is the probability of being "treated" (smoking) as a function of the potential confounders
- Fact: the distribution of X given p(X) is the same whether Z=1 or Z=0
  - The treated (smokers) and untreated (non-smokers) within a propensity score stratum are alike with respect to the covariates (age, gender, SES variables)
- Strategy:
  - Estimate the propensity score using logistic regression or other classification method
  - Stratify into quintiles of the estimated propensity score
  - Estimate the treatment effect within each stratum
  - Pool the estimates across strata

20

4. Survival Analysis

- Time to event observations when there is censoring:
  - Ungrouped data: using exact event times
  - Grouped data: using time intervals or bins (More to come!)
- Ungrouped data:
  - Calculate Kaplan-Meier estimates of S(t)
  - Use Cox regression model: assumes proportional hazards

21

4.1a Survival Analysis and Ungrouped Data

- Time to event data (observations) when there is censoring, using exact event times (ungrouped data)
- Hazard = h_t = instantaneous risk of event at time t = # events/# at risk
- Survivor function, S(t) = Pr(Survival beyond time t)
- Kaplan-Meier estimate of S(t)

$$S(t_i) = \prod_{i \le t}(1 - \frac{y_i}{n_i}) = \prod_{i \le t}(1 - h_i) = (1 - h_i)S(t_{i-1})$$

22

4.1b Survival Analysis and Ungrouped Data

- Weibull distribution provides a general form of the survivor function $S(t) = e^{-(\lambda t)^p}$
  - p = 1: Exponential distribution (constant hazard over time)
  - p > 1: Increasing hazard
  - p < 1: Decreasing hazard
- Complementary log-log transformation = CLL = log(-log S(t))
  - Plot of CLL vs log t should be a straight line if Weibull distribution fits
  - SE for CLL helps calculate confidence interval for S(t) within (0,1)

23

5.1a Interpretation of Cox Regression Coefficients

- $\log h(t, X) = \log h_0(t) + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_p X_p$
- Interpretation of $h_0(t)$*:*

  Hazard (incidence) rate as a function of time when all X's are zero
- β_1 is the change in the log hazard rate when X_1 increases one unit and the other X's remains fixed;

  or
- β_1 is the log hazard rate ratio associated with a one unit increase in X_1 and the other X's remains fixed.

24

5.1b Interpretation of Cox Regression Coefficients

- The model for the expected hazard rate is:

$$h(t;X)= h_0(t)\times e^{\beta_1 X_1}\times e^{\beta_2 X_2}\times\times\times\times e^{\beta_p X_p}$$

$$h(t;X)=h_0(t)\times e^{X\beta}$$

- And, e^{β} is the relative hazard associated with a one unit change in X_1 (i.e., X_1+1 vs. X_1), holding other X's constant, independent of time

Note: $e^{X\beta}$ "multiplies" the baseline hazard $\lambda_0(t)$ by the same amount regardless of the time t. This is therefore a "proportional hazards" model – the effect of any (fixed) X is the same at any time during follow-up

25

6. Inference from Regression Models

- Estimate β_j, SE(β_j)
  - 95% confidence intervals
  - Hypothesis tests, p-values
- Estimate linear combination of β_j's, SE(linear combination of β_j's)
  - 95% confidence intervals
  - Hypothesis tests, p-values
- Compare Extended versus Null Models
  - Test hypothesis that multiple β_j's equal 0

26

6.1a Inference for β_j in a MLR Model

- Use a **partial t-test** for a test of hypothesis about a specific β_j :
  H_0: $\beta_j = 0$
  H_a: $\beta_j \neq 0$ (two-sided test)
- Reject Ho at the $\alpha \cdot 100\%$ level if $\left|\frac{\hat{\beta}_j}{se_{\hat{\beta}_j}}\right| > t_{1-\alpha/2,df}$

where $\alpha = 0.05$, n = # observations, p = # Xs,
df = n-(p+1) = n-p-1

- 100(1- $\alpha \cdot$)% CI for β_j

$$\hat{\beta}_j \pm t_{1-\alpha/2,\,df} se_{\hat{\beta}_j}$$

27

6.1b Inference for β_j in a LR, Cox or LLR Model

- Use a **Wald test (Z test)** for a test of hypothesis about a specific β_j :
  H_0: $\beta_j = 0$
  H_a: $\beta_j \neq 0$ (two-sided test)
- Reject Ho at the $\alpha \cdot 100\%$ level if $\left|\frac{\hat{\beta}_j}{se_{\hat{\beta}_j}}\right| > Z_{1-\alpha/2}$

where $\alpha = 0.05$

- 100(1- $\alpha \cdot$)% CI for β_j

$$\hat{\beta}_j \pm Z_{1-\alpha/2} se_{\hat{\beta}_j}$$

28

6.2 Inference for $\beta_i+\beta_j$ in a Regression Model

- Use the **lincom** command in Stata for a test of hypothesis about a specific linear combination of β's:
  H_0: $\beta_i + \beta_j = 0$
  H_a: $\beta_i + \beta_j \neq 0$ (two-sided test)
- Reject Ho at the $\alpha \cdot 100\%$ level if $\left|\frac{\hat{\beta}_i + \hat{\beta}_j}{se_{\hat{\beta}_i+\hat{\beta}_j}}\right| > t_{1-\alpha/2,df}$
- Beware: $se_{\hat{\beta}_i+\hat{\beta}_j} \neq \sqrt{se^2_{\hat{\beta}_i} + se^2_{\hat{\beta}_j}}$
- 100(1- $\alpha \cdot$)% CI for $\beta_i + \beta_j$

$\hat{\beta}_i+\hat{\beta}_j \pm t_{1-\alpha/2,\,df} se_{\hat{\beta}_i+\hat{\beta}_j}$ *(or Z as appropriate)*

29

6.3a Testing H_o: $\beta_{p+1} = \beta_{p+2} \ldots = \beta_{p+s}=0$ in MLR

- Test of H_o: $\beta_{p+1} = \beta_{p+2} \ldots = \beta_{p+s}=0$ with

$$F=\frac{\frac{\sum_i (\hat{Y}_{N_i}-\hat{Y}_{E_i})^2}{s}}{\frac{\sum_i (Y_i-\hat{Y}_{E_i})^2}{(n-p-s-1)}} = \frac{\frac{\sum_i (\hat{Y}_{N_i}-\hat{Y}_{E_i})^2}{s}}{\hat{\sigma}^2_E}$$

- Under the null model ($\beta_{p+1} = \beta_{p+2} \ldots = \beta_{p+s}=0$), $F \sim F_{s,\,n-p-s-1}$ (F distribution with s and n-p-s-1 df)
- Use **test** command in Stata

30

6.3b Testing H_o: $\beta_{p+1} = \beta_{p+2} \ldots = \beta_{p+s} = 0$ in LR, Cox or LLR

- Likelihood Ratio Test (LRT) statistic for comparing nested models is -2 times the difference between the log likelihoods (LLs) for the Null -vs- Extended models – the Δ obtained **is identical** to Δ from an analysis of variance test for linear regression models
- **- 2** Δ LL $\sim \chi^2_s$
- Use the **lrtest** command in combination with fitting the two models with the logistic command and storing the estimation results (**est store** command)

31

6.4 Confounding

- In epidemiological terms, Z is a "confounder" of the relationship of Y with X if Z is related to both X and Y and Z is not is the causal pathway between X and Y
- In statistical terms, Z is a "confounder" of the relationship of Y with X if it is not in the causal pathway X $\rightarrow$Y and the X coefficient substantially changes when Z is added to a regression of Y on X
- For example, consider the two models
  $Y = \beta_0 + \beta_1 X + \varepsilon_1$
  $Y = \alpha_0 + \alpha_1 X + \alpha_2 Z + \varepsilon_2$
  then Z is a confounder of the X,Y relationship if $\alpha_1 \neq \beta_1$
  Guideline: coefficient changes by > 15%
- This definition of confounding applies to all regression models: MLR, LR, Cox and LLR

32

6.5 Effect Modification

- Effect modification (interaction) is present when the coefficient for an *X* variable differs depending on the values of one or more other *X*s (e.g. the relationship between X and the outcome varies by the level of one or more other variables)
- This concept applies to all the regression models: MLR, LR, Cox and LLR

33

6.6 Example 1: Effect Modification

- For illustration consider two categorical (dichotomous) predictors (smoking and gender)
  - X_1 = smoker where 0= non-smoker, 1 = smoker
  - X_2 = gender where 0 = male, 1=female
- The linear predictor in regression models, including the interaction term, is:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \cdot X_2$$

34

6.6a Regression Coefficients for Outcome as a Function of Gender and Smoking History

	X_1=0 (No)	X_1=1 (Yes)
X_2=0 (Male)	β_0	$\beta_0+\beta_1$
X_2=1 (Female)	$\beta_0+\beta_2$	$\beta_0+\beta_1+\beta_2+\beta_3$

35

6.6b Interpretations of Coefficients

	MLR	LR	Cox PH
β_0	Ave *Y* for non-smoking males	log odds for non-smoking males	No β_0 in model -- model uses baseline hazard $\lambda_0(t)$
β_1	Ave diff in *Y* smokers vs non-smokers among males	log odds ratio smokers vs non-smokers among males	log relative hazard smokers vs non-smokers among males
β_2	Ave diff in *Y* females vs males among non-smokers	log odds ratio females vs males among non-smokers	log relative hazard females vs males among non-smokers

36

6.6c Interpretations of Coefficients

	MLR	LR	Cox PH
$\beta_1+\beta_3$	Ave diff in Y smokers vs non-smokers among females	log odds ratio smokers vs non-smokers among females	log relative hazard smokers vs non-smokers among females
$\beta_2+\beta_3$	Ave diff in Y females vs males among smokers	log odds ratio females vs males among smokers	log relative hazard females vs males among smokers

37

6.6d Interpretations of Coefficients

	MLR	LR	Cox PH
β_3	Difference in ave diff in Y smokers vs non-smokers for females vs males	Difference in log odds ratio smokers vs non-smokers for females vs males	Difference in log hazard ratio smokers vs non-smokers for females vs males
β_3	Difference in ave diff in Y females vs males for smokers vs non-smokers	Difference in log odds ratio females vs males for smokers vs non-smokers	Difference in log hazard ratio females vs males for smokers vs non-smokers

38

6.7 Example 2: Effect Modification

- For another illustration, consider a categorical predictor (smoking) and a continuous predictor (systolic BP):
  - X_1 = smoker where 0= non-smoker, 1 = smoker
  - X_2 = systolic BP (mm Hg) centered at *140*
- The linear predictor in regression models, including the interaction term $X_3 = X_1 \cdot (X_2 - 140)$ is:

$$\beta_0 + \beta_1 X_1 + \beta_2 (X_2 - 140) + \beta_3 X_3$$

39

6.7a Interpretations of Coefficients

	MLR	LR	Cox PH
β_0	Ave *Y* for non-smokers with 140 BP	log odds for non-smokers with 140 BP	No β_0 in model -- model uses baseline hazard $\lambda_0(t)$
β_1	Ave diff in *Y* smokers vs non-smokers at BP of 140	log odds ratio smokers vs non-smokers at BP of 140	log relative hazard smokers vs non-smokers at BP of 140
β_2	Ave diff in *Y* per unit increased BP for non-smokers	log odds ratio per unit increased BP for non-smokers	log relative hazard per unit increased BP for non-smokers

40

6.7b Interpretations of Coefficients

	MLR	LR	Cox PH
$\beta_2+\beta_3$	Ave diff in *Y* per unit increased BP for smokers	log odds ratio per unit increased BP for smokers	log relative hazard per unit increased BP for smokers
β_3	Difference in ave diff in *Y* per unit increased BP smokers vs non-smokers	Difference in log odds ratio per unit increased BP smokers vs non-smokers	Difference in log hazard ratio per unit increased BP smokers vs non-smokers

6.8 Example 3: Effect Modification and Spline for BP

- For another illustration, consider a categorical predictor (smoking) and a continuous predictor (systolic BP):
  - X_1 = smoker where 0= non-smoker, 1 = smoker
  - X_2 = systolic BP (mm Hg) centered at *140*
  - $X_3 = (X_2 - 140)^+$ Spline at 140 mm BP
- The linear predictor in regression models, including the interaction terms $X_4 = X_1(X_2-140)$ and $X_5 = X_1(X_2-140)^+$ is:

$$\beta_0 + \beta_1 X_1 + \beta_2(X_2 - 140) + \beta_3 X_3 + \beta_4 X_1(X_2 - 140) + \beta_5 X_1(X_2 - 140)^+$$

6.8a Interpretations of Coefficients

	MLR	LR	Cox PH
β_0	Ave *Y* for non-smokers with 140 BP	log odds for non-smokers with 140 BP	No β_0 in model -- model uses baseline hazard $\lambda_0(t)$
β_1	Ave diff in *Y* smokers vs non-smokers at BP of 140	log odds ratio smokers vs non-smokers at BP of 140	log relative hazard smokers vs non-smokers at BP of 140
β_2	Ave diff in *Y* per unit increased BP for non-smokers at or before the breakpoint	log odds ratio per unit increased BP for non-smokers at or before the breakpoint	log relative hazard per unit increased BP for non-smokers at or before the breakpoint

6.8b Interpretations of Coefficients

	MLR	LR	Cox PH
β_3	Difference in ave diff in *Y* per unit increased BP in non-smokers after vs before the breakpoint	Difference in log odds ration per unit increased BP in non-smokers after vs before the breakpoint	Difference in log hazard ratio per unit increased BP in non-smokers after vs before the breakpoint
$\beta_2+\beta_4$	Ave diff in *Y* per unit increased BP for smokers at or before the breakpoint	log odds ratio per unit increased BP for smokers at or before the breakpoint	log relative hazard per unit increased BP for smokers at or before the breakpoint

6.8c Interpretations of Coefficients

	MLR	LR	Cox PH
β_4	Difference in ave diff in Y per unit increased BP at or before the breakpoint smokers vs non-smokers	Difference in log odds ratio per unit increased BP at or before the breakpoint smokers vs non-smokers	Difference in log hazard ratio per unit increased BP at or before the breakpoint smokers vs non-smokers

45

6.8d Interpretations of Coefficients

	MLR	LR	Cox PH
$\beta_2+\beta_3+\beta_4+\beta_5$	Ave diff in Y per unit increased BP for smokers after the breakpoint	log odds ratio per unit increased BP for smokers after the breakpoint	log relative hazard per unit increased BP for smokers after the breakpoint
β_5	Difference in ave diff in Y per unit increased BP after vs before the breakpoint for smokers vs non-smokers	Difference in log odds ratio per unit increased BP after vs before the breakpoint for smokers vs non-smokers	Difference in log relative hazard per unit increased BP after vs before the breakpoint for smokers vs non-smokers

46

6.9a Selecting a Regression Model

- Selecting models depends on:
  - Question of interest
  - Purpose (e.g. etiology, adjustment, prediction, differing costs for measuring Xs)
- Check model fit:
  - MLR: Inspect residuals plots, adjusted variables plots (AVPs)
    - Look for non-linear patterns, influential points, and changing variance
    - Check influence of influential points by using dfits or other measures
  - LR: Inspect plots of observed values versus predicted values, Hosmer-Lemeshow goodness of fit test
    - Look for patterns, influential points, and changing variance
    - Check influence of influential points
  - Cox and LLR: Inspect complementary log-log plots[47]

6.9b Selecting a Regression Model

- Criteria used:
  - cross-validated measures of prediction error
  - AIC = -2(model Log Likelihood) + 2(#parameters) for MLR,LR, Cox and LLR
  - R^2 is not a good measure
- More!

48

7.a Summary

- Sample size estimation is an important component of study design and influences statistical analysis.
- Generalized linear models and regression analysis is a statistical method for describing a “response” or “outcome” variable as a simple function of “explanatory” or “predictor” variables (X)
- Regression analysis includes model selection and methods for checking model fit

49

7.b Summary

- Propensity scores can provide a method for controlling for possible confounders.
- Survival analysis is used for time to binary event data, especially in the presence of censored observations.
  - Kaplan-Meier estimates of the survivor function
  - Cox proportional hazards regression model for which estimated regression coefficients are log hazard ratios.

50