# AS.280.347
# CLASS 1.3

- Look at data displays & logistic regression models!
- Propensity scores
- Using R for propensity scores

# Module 1: Smoking and risk of disease

- Question 1.1 (Q1.1): How does the risk of disease compare for smokers and otherwise similar non-smokers?
- Question 1.2 (Q1.2): Does the contribution of smoking to the risk of disease vary by sex or SES?
- To address each question, we want:
  - a data display
  - a statistical analysis
- We will answer these questions using data from the National Medical Expenditures Survey (NMES)

# Today's agenda

- Group presentations and critiques Q1.1:

  *How does the risk of disease compare for smokers and otherwise similar non-smokers?*

  - Updated data displays
  - Logistic regression models

- Plans to improve displays/models

- Propensity score approach

# Confounding

- Goal is to estimate the effect of a “treatment” or “risk factor” (e.g., ever smoking) on an outcome (e.g., major smoking-caused disease) by *comparing otherwise similar persons with and without the risk factor.*

- How could we account for any possible confounding variables in a logistic regression analysis?

# Review of logistic regression

- How could we account for any possible confounding variables in a logistic regression analysis?
  - We could include potential confounding variables as covariates in our analysis using multivariable logistic regression:

$$\log(odds\ of\ death) = \beta_0 + \beta_1 \cdot (gestational\ age) + \beta_2 \cdot twin$$

  - We interpret the regression coefficients in a multivariable model as ***ceteris paribus*** – holding all other things equal
  - $\beta_1 = \log(OR)$ for a one-unit change in gestational age, **holding twin status constant**
  - $\beta_2 = \log(OR)$ comparing twins to singleton births, **holding gestational age constant**

# Scenario for today's discussion

- Health response: Y
  - Major smoking caused disease
- Binary treatment or risk factor: Z=1,0
  - Ever smoker
- Potential confounders: X
  - Age
  - Gender
  - SES: poverty, education
  - Marital status
  - Etc

# Stratification to account for confounding

- Stratify by the covariate
- Estimate the difference in mean outcome within each covariate stratum
- Pool the stratum-specific values

# Example: Stratifying by income

| Income Level | Probability of MSCD (n) | | Smoking effect Log OR | Std error | W=1/var | W* Log OR |
|---|---|---|---|---|---|---|
| | Ever smokers | Never smokers | | | | |
| 1 (Poverty) | .158 (677) | .079 (579) | | | | |
| 2 | .191 (303) | .120 (292) | | | | |
| 3 | .187 (925) | .092 (739) | | | | |
| 4 | .131 (2083) | .094 (1548) | | | | |
| 5 | .105 (2607) | .073 (1892) | | | | |
| Pooled | | | | | | |

# Example: Stratifying by income

| Income Level | Probability of MSCD (n) | | Smoking effect Log OR | Std error | W=1/var | W* Log OR |
|---|---|---|---|---|---|---|
| | Ever smokers | Never smokers | | | | |
| 1 (Poverty) | .158 (677) | .079 (579) | **.777** | **.186** | | |
| 2 | .191 (303) | .120 (292) | **.553** | **.232** | | |
| 3 | .187 (925) | .092 (739) | **.820** | **.153** | | |
| 4 | .131 (2083) | .094 (1548) | **.370** | **.109** | | |
| 5 | .105 (2607) | .073 (1892) | **.397** | **.109** | | |
| Pooled | | | | | | |

# Example: Stratifying by income

| Income Level | Probability of MSCD (n) | | Smoking effect Log OR | Std error | W=1/var | W* Log OR |
|---|---|---|---|---|---|---|
| | Ever smokers | Never smokers | | | | |
| 1 (Poverty) | .158 (677) | .079 (579) | .777 | .186 | **.112** | **.087** |
| 2 | .191 (303) | .120 (292) | .553 | .232 | **.072** | **.040** |
| 3 | .187 (925) | .092 (739) | .820 | .153 | **.165** | **.135** |
| 4 | .131 (2083) | .094 (1548) | .370 | .109 | **.326** | **.121** |
| 5 | .105 (2607) | .073 (1892) | .397 | .109 | **.326** | **.129** |
| Pooled | | | | | **1** | **.512** |

# Example: MSCD

- Regression of Y on X and indicator variables of the strata is identical to weighting the log ORs inversely related to their variances.

```
> model6 = glm(mscd ~ eversmk + as.factor(income), family=binomial(link="logit"), data=nmesData)
> summary(model6)$coefficients
                     Estimate Std. Error    z value      Pr(>|z|)
(Intercept)       -2.27864855 0.09510615 -23.959004 7.444629e-127
eversmk            0.51664113 0.06211029   8.318124  8.936644e-17
as.factor(income)2 0.30670165 0.14266039   2.149873  3.156530e-02
as.factor(income)3 0.19255875 0.11127643   1.730454  8.354917e-02
as.factor(income)4 -0.07918846 0.10105222  -0.783639  4.332520e-01
as.factor(income)5 -0.34393439 0.10095077  -3.406952  6.569275e-04
```

- A faster method of pooling the evidence!

# What to do with many potential confounders?

- 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 fore the two treatment groups within each stratum of the variable
  - One such variable is the **propensity score**

# What is a 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)

# Propensity score strategy – idea

- Estimate the propensity score using logistic regression (or other classification method)
  - Estimate probability of being a smoker, given age, sex, SES, etc
  - Estimate Pr(eversmk = 1 | age, sex, SES, etc)
- Stratify by this propensity score (perhaps into 5 groups based on the quintiles of the scores)
- Estimate the treatment effect within each stratum
  - Calculate the log OR (or OR) of MSCD, comparing smokers to non-smokers, within each PS group
- Pool the estimates across strata
  - Use inverse-variance weighting to combine estimates

# Propensity score strategy – implementation

- Estimate the propensity score using logistic regression (or other classification method)

```
propModel <- glm( eversmk ~ ??? , family=binomial(link="logit"))
predLogOdds <- predict(propModel)
predProb <- exp(predLogOdds)/(1+exp(predLogOdds))
propScores <- predProb
```

- Stratify by this propensity score (perhaps into 5 groups based on the quintiles of the scores)

```
psCutoffs <- quantile(propScores, probs=c(0,0.25, 0.5, 0.75, 1))
ps.groups <- cut(propScores, psCutoffs, include.lowest=TRUE)
```

- Estimate the treatment effect within each stratum and pool the estimates across strata

```
glm( mscd ~ eversmk + ps.groups , family=binomial(link="logit"))
```

# Pros and cons of propensity scores

- Organizes the analysis into 2 steps
  1. Estimate the probability of treatment given the covariates
  2. Compare treatment groups within strata of this probability

- Easy to picture the evidence for the **binary treatment** effect
  - Most natural with binary treatment (extensions possible, but awkward)

- Not as simple to study effect modifications (interactions)

- No method controls for unmeasured confounders, regardless of what is claimed

# Assignment 1.3

- Improve your data display to answer Q1.1: ***How does the risk of disease compare for smokers and otherwise similar non-smokers?***
- Update your logistic regression model to answer Q1.1. What does this model say about Q1.1? *Be sure to focus on answering the question being asked!*
- Estimate propensity scores for the treatment of smoking (eversmk); that is, use logistic regression to estimate the probability of smoking given possible confounders.
- Use logistic regression with quintiles of your propensity scores to answer Q1.1. Interpret the results.

- Work together in groups!
- Submit your assignment in R markdown through Blackboard by Sunday @ midnight.

# Thinking ahead…

- What statistical analysis will effectively use the NMES data to address Q1.1?
  - Multiple logistic regression
  - Propensity scores

- What displays and statistical analysis will address Question 1.2 (Q1.2):
  - *Does the contribution of smoking to the risk of disease vary by sex or SES?*

# References for propensity scores


- Rosenbaum and Rubin, 1983. *Biometrika*, 70: 41-55.
- Rubin. 1997. *Annals of Internal Medicine*, 127: 757-763.