AS.280.347
CLASS 1.2

- Look at data displays!
- Review of logistic regression

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?

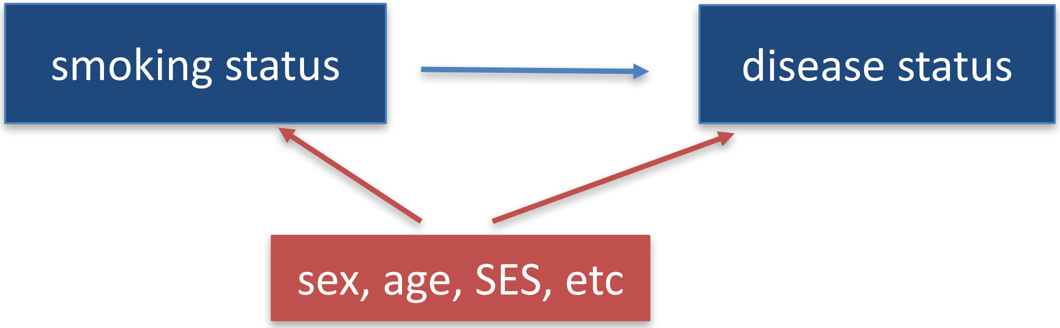
- Question 1.2 (Q1.2): Does the contribution of smoking to the risk of disease vary by sex or SES?

- To address each question, we want:
  - a data display
  - a statistical analysis

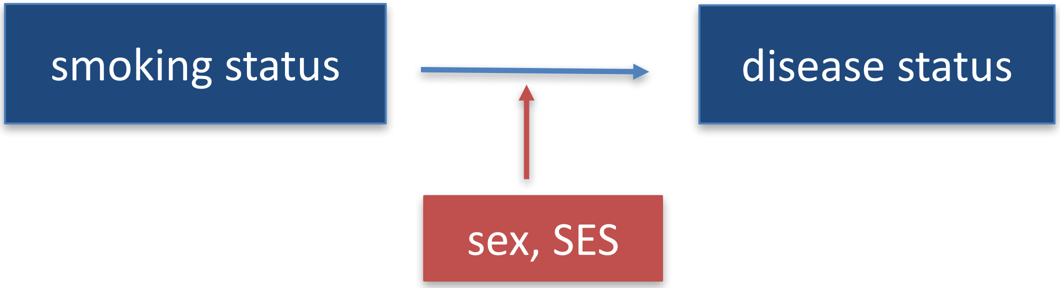
- We will answer these questions using data from the National Medical Expenditures Survey (NMES)

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?



Today's agenda

- Group discussion and critiques of NMES data displays to address Q1.1:
  - *How does the risk of disease compare for smokers and otherwise similar non-smokers?*
- Plans to improve displays
- Review of logistic regression

Questions for discussion

- What are the characteristics of effective displays?
- How can the current displays that address Q1.1 be improved?
- How can the process of working together be improved?

- What statistical analysis will effectively use the NMES data to address Q1.1?
  - Multivariable logistic regression
  - Propensity scores (next week)

- 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?*

Review of logistic regression

In Public Health Biostatistics we used logistic regression to estimate the risk of infant mortality as a function of gestational age, parity and other factors

- Used for **binary** outcome variables
  - Ex: infant mortality (1=infant died, 0=infant survived)

- Models the log odds of the outcome:
  - $\log(odds\ of\ Y = 1) = \beta_0 + \beta_1 X$
  - Ex: $\log(odds\ of\ death) = \beta_0 + \beta_1 \cdot (gestational\ age)$

- We transform to get probability/risk:
  - $\log(odds\ of\ Y = 1) = \beta_0 + \beta_1 X$
  - $(odds\ of\ Y = 1) = e^{\beta_0 + \beta_1 X}$
  - $P(Y = 1) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$

Review of logistic regression

- Is the mortality risk (or odds) higher for twins than singleton births?

	Singleton	Twin	Total
Survived	8899	187	9086
Died	526	71	597
Total	9425	258	9683

- Odds of death for twins: **71/187 = .38**
- Odds of death for singletons: **526/8899 = .059**
- Odds ratio of death for twins as compared to singletons:

  OR = (odds for twins)/(odds for sing) = (71/187)/(526/8899) = 6.42

- Log odds ratio: **$\log_e$(OR) = log(6.42) =1.86**

Review of logistic regression

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

$$twin = \begin{cases} 1 & if\ twin \\ 0 & if\ singleton \end{cases}$$

```
> model1 = glm(death ~ twins, family=binomial(link="logit”))
> summary(model1)
```

```
Coefficients:
              Estimate  Std. Error z value  Pr(>|z|)
(Intercept)   -2.82839    0.04487  -63.03   <2e-16 ***
twins          1.85996    0.14644   12.70   <2e-16 ***
---
Null deviance: 4483.1  on 9682  degrees of freedom
Residual deviance: 4361.6  on 9681  degrees of freedom
AIC: 4365.6
Number of Fisher Scoring iterations: 5
```

Review of logistic regression

```
> model1 = glm(death ~ twins, family=binomial(link="logit”))
> summary(model1)
```

β | $SE(\beta)$ | test statistic | *p*-value

```
Coefficients:
              Estimate  Std. Error z value  Pr(>|z|)
(Intercept)   -2.82839    0.04487  -63.03   <2e-16 ***
twins          1.85996    0.14644   12.70   <2e-16 ***
---
Null deviance: 4483.1  on 9682  degrees of freedom
Residual deviance: 4361.6  on 9681  degrees of freedom
AIC: 4365.6
Number of Fisher Scoring iterations: 5
```

$1.86 = \beta_1 = log(OR)$

The log odds of death, comparing twins to singleton births, is 1.86.

$6.42 = e^{1.86} = e^{\beta_1} = OR$

The odds of death for twins is 6.42 times the odds of death for singleton births.

Review of logistic regression

- Does the odds of death decrease with increasing gestational age?

$$\log(\boldsymbol{odds\ of\ death}) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} \cdot (\boldsymbol{gestational\ age})$$

- $\log(odds\ of\ death \mid ga = 41\ weeks) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} \cdot (\mathbf{41})$
- $\log(odds\ of\ death \mid ga = 40\ weeks) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} \cdot (\mathbf{40})$

- Difference:

$$\log(odds \mid ga = 41) - \log(odds \mid ga = 40)$$

$$= (\beta_0 + 41\beta_1) - (\beta_0 + 40\beta_1) = \boldsymbol{\beta_1}$$

- Log odds ratio:

$$\log(\mathrm{OR}) = \log\left(\frac{odds \mid ga=41}{odds \mid ga=40}\right)$$

$$= \log(odds \mid ga = 41) - \log(odds \mid ga = 40) = \boldsymbol{\beta_1}$$

- Odds ratio: $OR = e^{\log(OR)} = \boldsymbol{e^{\beta_1}}$

Review of logistic regression

```
> model2 = glm(death ~ gestage, family=binomial(link="logit”))
> summary(model2)
```

β

```
Coefficients:
             Estimate Std. Error  z value   Pr(>|z|)
(Intercept)    2.3274     0.3943    5.902   3.59e-09 ***
gestage       -0.1359     0.0108  -12.584   < 2e-16 ***

Null deviance:    4483.1  on 9682  degrees of freedom
Residual deviance: 4328.0  on 9681  degrees of freedom
AIC: 4332
```

$-0.1359 = \beta_1 = log(OR)$

An additional week of gestational age is associated with a decrease of .14 in the log odds of death.

$0.87 = e^{-.1359} = e^{\beta_1} = OR$

An additional week of gestational age is associated with a 13% decrease in the odds of infant death.

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:

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

  - 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**

Assignment 1.2

- **Q1.1: *How does the risk of disease compare for smokers and otherwise similar non-smokers?***
  - Improve your data display to answer this question.
  - Fit a logistic regression model to answer this question. Interpret your coefficients and significance tests to answer the question: what does this model say about Q1.1?
  - Work together in groups!
  - Submit your display in R markdown through Blackboard by Sunday @ midnight.

- Consider completing any/all of these available DataCamp modules:
  - Introduction to R
  - Reporting with R Markdown: Authoring R Markdown Reports
  - Reporting with R Markdown: Embedding Code
  - Reporting with R Markdown: Compiling Reports
  - Introduction to the Tidyverse

- Next week in class we will again start with presentations/critiques of your displays.