Biostatistics 208

Lecture 1: Overview & Linear Regression Intro.

Steve Shiboski
Division of Biostatistics, UCSF

January 7, 2014

Organization

- Office hours by appointment
  (CBL 5730)
- E-mail to make an appointment or get questions answered quickly:
  steve@biostat.ucsf.edu
- Download lecture slides, labs, data & homework from course site

Textbook - VGSM, 2nd edition

Statistics for Biology and Health

Eric Vittinghoff · David Glidden · Steve Shiboski · Charles E. McCulloch

Regression Methods in Biostatistics

Linear, Logistic, Survival, and Repeated Measures Models, *Second Edition*

This new edition provides a unified, in-depth, readable introduction to the multipredictor regression methods most widely used in biostatistics: linear models for continuous outcomes, logistic models for binary outcomes, the Cox model for right-censored survival times, repeated-measures models for longitudinal and hierarchical outcomes, and generalized linear models for counts and other outcomes.

Treating these topics together takes advantage of all they have in common. The authors point out the many-shared elements in the methods they present for selecting, estimating, checking, and interpreting each of these models. They also show that these regression methods deal with confounding, mediation, and interaction of causal effects in essentially the same way.

The examples, analyzed using Stata, are drawn from the biomedical context but generalize to other areas of application. While a first course in statistics is assumed, a chapter reviewing basic statistical methods is included. Some advanced topics are covered but the presentation remains intuitive. A brief introduction to regression analysis of complex surveys and notes for further reading are provided. For many students and researchers learning to use these methods, this one book may be all they need to conduct and interpret multipredictor regression analyses.

In the second edition of *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*, the authors have substantially revised and expanded the core chapters of the first edition, and added two new chapters.

From the reviews of the first edition:

"This book provides a unified introduction to the regression methods listed in the title...The methods are well illustrated by data drawn from medical studies... A real strength of this book is the careful discussion of issues common to all of the multipredictor methods covered."
Journal of Biopharmaceutical Statistics, 2005

"This book is not just for biostatisticians. It is, in fact, a very good, and relatively nonmathematical, overview of multipredictor regression models. Although the examples are biologically oriented, they are generally easy to understand and follow...I heartily recommend the book"
Technometrics, February 2006

"Overall, the text provides an overview of regression methods that is particularly strong in its breadth of coverage and emphasis on insight in place of mathematical detail. As intended, this well-unified approach should appeal to students who learn conceptually and verbally."
Journal of the American Statistical Association, March 2006

Statistics / Life Sciences, Medicine, Health Sciences

ISBN 978-1-4614-1352-3

9 781461 413523

▸ springer.com

SBH

Vittinghoff · Glidden
Shiboski · McCulloch

Regression Methods in Biostatistics

2nd Ed.

Statistics for Biology and Health

Eric Vittinghoff
David Glidden
Steve Shiboski
Charles E. McCulloch

Regression Methods in Biostatistics

Linear, Logistic, Survival, and Repeated Measures Models

Second Edition

Springer

Lectures

- Linear regression for continuous outcomes

  7 lectures on simple and multiple linear regression

- Logistic regression for binary outcomes

  4 lectures on logistic regression and alternatives

Labs

- Thursdays in CBL 6702 & 6704
- 10:30-12:00
- Please be on time
- Labs show how to use Stata to implement methods discussed in class
- We supply the data and commands, plus an interpretive handout will be posted on the course site at the end of the lab

Homework

- 70% of grade
- Four homework assignments; *missing one can making passing difficult*
- Your responsibility to meet deadlines; late home works not accepted; *if you are going to be out of town, please make arrangements*
- Assignments posted on the course web site (first one will be posted today)
- Handed back the next week with comments and discussion session

Heads up: Biostat 209 Project

This will be a written report of a data analysis, similar to a brief, focused research manuscript but with more statistical detail than usual.

Goal: Address a single substantive issue or a few closely related issues; this should not be a comprehensive report of all findings from a study.

- Students taking 209 next quarter should start thinking about the project now
- Identify a candidate dataset, ideally linked to your own research interests
- The project can be the basis for a submitted abstract/manuscript

Review of Exploratory Data Analysis

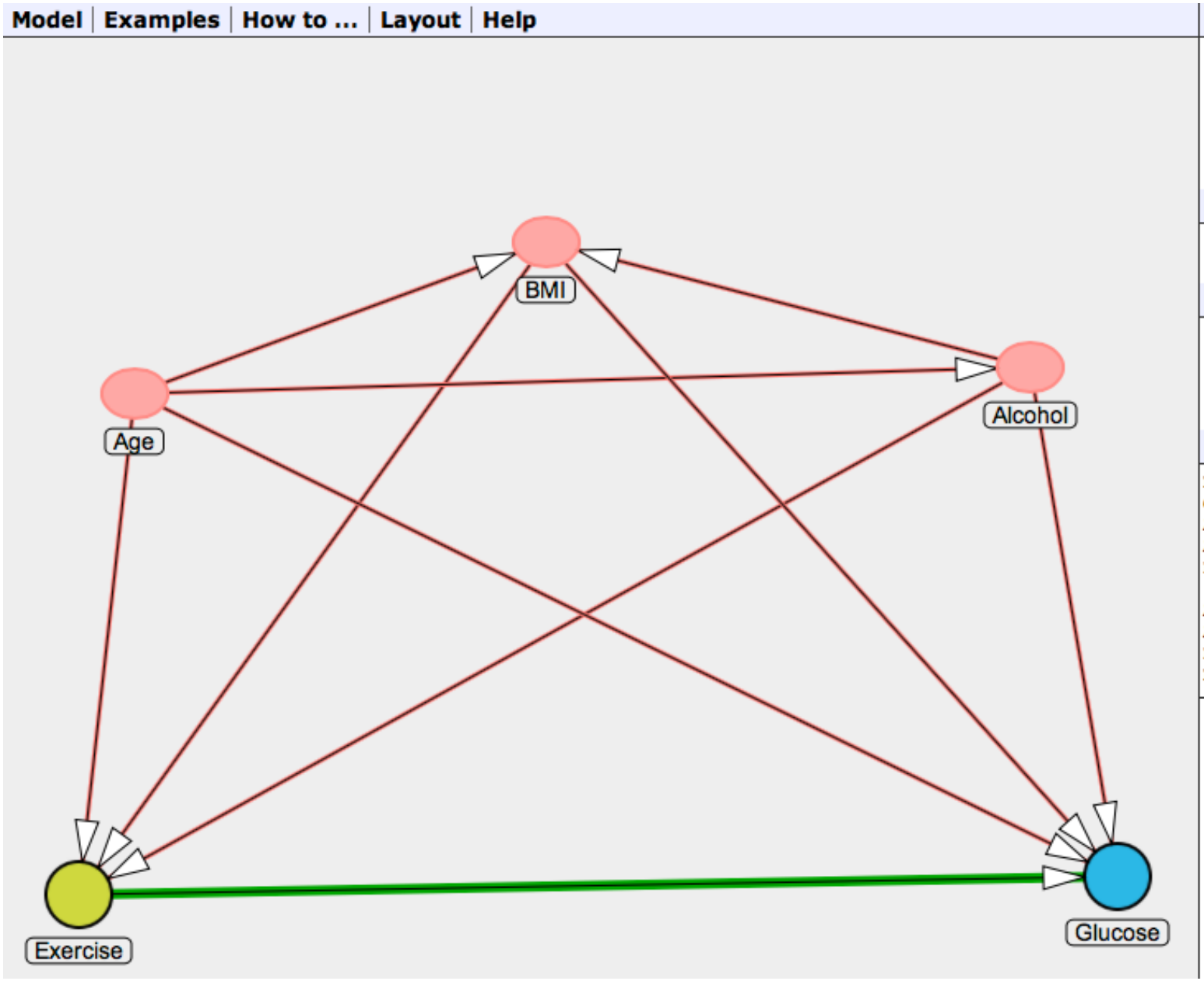
- A key part of regression modeling is exploratory data analysis of key variables (as covered in Biostat 200)
- Exploratory assessments should be made prior to any data analysis
- The first assignment & lab review some of these concepts, and a few slides covering key topics are also included at the end of the pdf version of these slides posted on the course web site

Regression models for the relationship between outcomes and predictor variables

- Research questions frequently focus on the relationship between an outcome and additional *predictor variables*
- For single categorical predictors (e.g. treatment assignment in randomized trials), simple analysis methods (e.g. *t*-tests & chi squared tests) usually suffice
- For multiple predictors, and/or continuous predictors, the relationship with the outcome is frequently very complex
- Statistical models are useful tools for summarizing relationships and making inferences when continuous and/or multiple predictors are involved

Example: relationship between exercise & glucose in HERS study (section 4.1)

- Sample DAG (from dagitty.net) suggests adjustment for age, alcohol & BMI



Multiple Predictor Regression Models

- Additive, linear statistical models for the relationship between an outcome and multiple predictor variables (aka "independent", "risk" or "exposure" variables)
- Outcome types include
  - binary (or categorical) indicators (e.g. *logistic models*)
  - continuous measures (e.g. *linear models*)
  - event times (e.g. *proportional hazards models*)
- Methods can be applied to dependent outcomes (*repeated measures*) and a range of study designs (*surveys, case-control*)
- Because regression models make strong assumptions about the relationship between outcome and predictors, assessing validity of assumptions is a critical part of any application

Common Goals in Regression Modeling

- **Evaluating a predictor of primary interest**
  - additional predictors selected both for face validity and on statistical grounds, guided by appropriate DAG(s)
  - focus on confounding, mediation and interaction
  - causal inference
- **Identifying multiple important predictors**
  - variables selected as indicated above
  - addressing issues of confounding, mediation and interaction is complicated; over-fitting & false-positive results become concerns
- **Outcome prediction / risk stratification**
  - focus on minimizing *prediction error*

Contrasting Multiple Regression Approaches

- **Similarities**

- Structural:

  - Predictors linked to the some feature of the outcome in an additive, linear fashion:
  - *Example*: linear regression
  - The conditional mean the outcome (y) is a linear function of the predictors (x_1, x_2, …, x_p):

$$\mathrm{E}\left[y|x_1, x_2, \cdots, x_p\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

  - *Example*: logistic regression
  - The conditional odds of outcome occurrence is a log-linear function of the predictors

$$\log\left[\frac{P(x_1,\cdots,x_p)}{1-P(x_1,\cdots,x_p)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Contrasting Multiple Regression Approaches

- **Similarities**
  - Operational (methods for using regression in practice are similar for different outcome types)
    - model construction & selection
    - statistical inference
    - model evaluation & assessment

Contrasting Multiple Regression Approaches

- **Differences**
  - Differing outcome types require different distributional assumptions
  - Interpretation of coefficients and predicted quantities differ between approaches
  - Alternative techniques for estimation are involved

Review: linear regression with a single predictor

- Linear model for the relationship between a continuous outcome and categorical/continuous predictor variable

Example: The HERS Study

- **H**eart and **E**strogen/Progestin **R**eplacement **S**tudy
- Randomized trial of Estrogen + Progestin for secondary prevention of coronary heart disease (CHD) events in postmenopausal women with existing CHD (Hulley S, *et al*; *JAMA* , 1998)
- 2763 women, ages 46 - 79, followed for an average 4.1 years
- Examples for today based on a random sample of 221 women at study entry
- Focus on the dependence of HDL cholesterol on smoking and waist circumference

Data description from Stata:

```
. describe

Contains data from ~/Documents/teaching/c2012/atcr/lecture2/data/hers-sample.dta
  obs:           221
 vars:             3                          18 Jan 2005 08:36
 size:         2,431 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
smoking         byte   %9.0g       noyes      current smoker
waist           float  %9.0g                  waist circumference (cm)
hdl             int    %9.0g                  hdl cholesterol (mg/dl)
-------------------------------------------------------------------------------
Sorted by:
```

Data listing

```
. list hdl smoking waist
     +----------------------+
     | hdl   smoking   waist |
     |----------------------|
  1. |  62         no    85.3 |
  2. |  46         no   106.7 |
  3. |  29         no      95 |
  4. |  52         no    98.2 |
  5. |  49        yes   100.7 |
     |----------------------|
  6. |  43         no    88.8 |
  7. |  63        yes      88 |
  8. |  45         no     102 |
  9. |  40         no    85.6 |
 10. |  53         no    71.6 |
     |----------------------|
 11. |  49         no    86.6 |
 12. |  72         no    74.3 |
 13. |  58         no      86 |
 14. |  51         no   109.5 |
 15. |  57         no    85.1 |
     |----------------------|
 16. |  52         no     139 |
 17. |  36         no      84 |
 18. |  43         no     119 |
 19. |  36         no    94.5 |
 20. |  53        yes     104 |
     |----------------------|
 21. |  54         no    76.7 |
 22. |  40         no    76.5 |
--more--
```

Listing more info about variables and their characteristics

```
. codebook
-------------------------------------------------------------------------------
smoking                                                  current smoker
-------------------------------------------------------------------------------
                  type:  numeric (byte)
                 label:  noyes
                 range:  [0,1]                        units:  1
         unique values:  2                    missing .:  0/221
            tabulation:  Freq.   Numeric  Label
                           194         0  no
                            27         1  yes
-------------------------------------------------------------------------------
waist                                            waist circumference (cm)
-------------------------------------------------------------------------------
                  type:  numeric (float)
                 range:  [57.5,139]                   units:  .1
         unique values:  145                  missing .:  0/221
                  mean:   92.5652
             std. dev:   14.5148
           percentiles:        10%       25%       50%       75%       90%
                                74        82      92.2       102       110
-------------------------------------------------------------------------------
hdl                                              hdl cholesterol (mg/dl)
-------------------------------------------------------------------------------
                  type:  numeric (int)
                 range:  [24,112]                     units:  1
         unique values:  56                   missing .:  2/221
                  mean:   50.2785
             std. dev:   13.8306
           percentiles:        10%       25%       50%       75%       90%
                                36        41        48        57        68
```

Two indiv. missing HDL

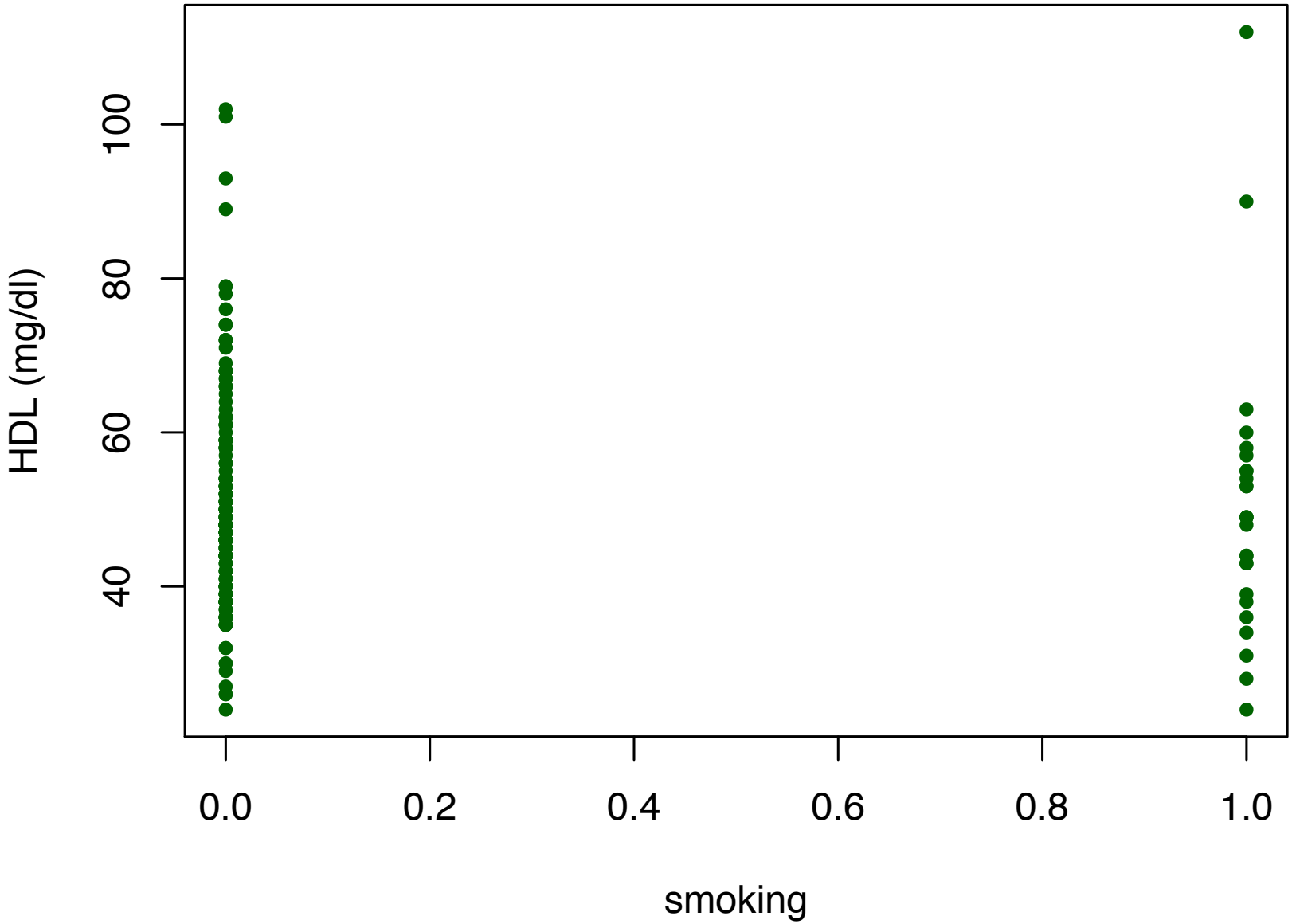
Summaries of included variables

```
. summarize hdl

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         hdl |        219    50.27854    13.83056        24        112

. summarize waist

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       waist |        221    92.56516    14.51481      57.5        139

. tab smoking

    current |
     smoker |      Freq.     Percent        Cum.
------------+-----------------------------------
         no |        194       87.78       87.78
        yes |         27       12.22      100.00
------------+-----------------------------------
      Total |        221      100.00
```

Example analyses using HERS data:

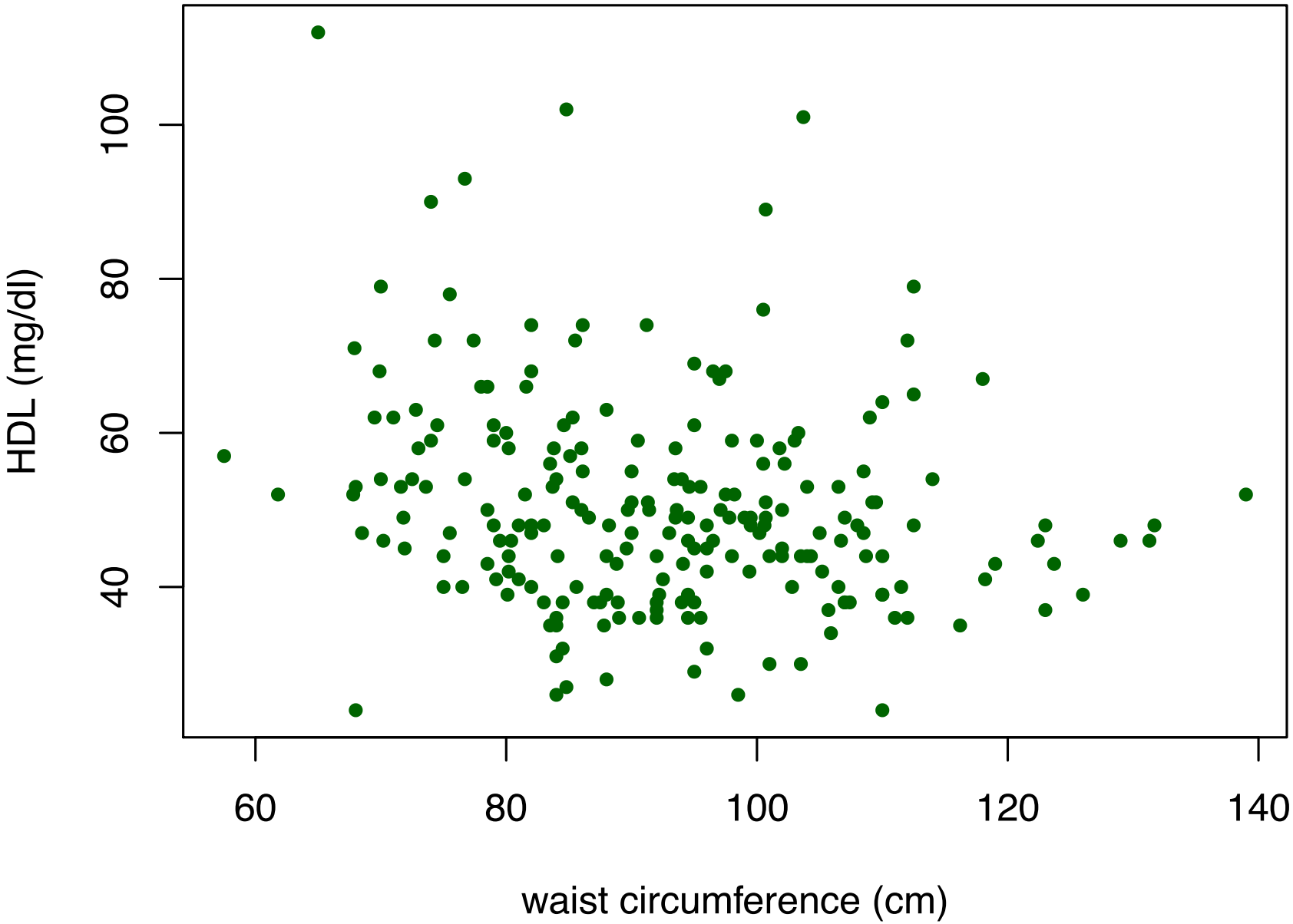
- Relationship between HDL cholesterol (*mg/dl*) and a single binary predictor representing smoking status (*Yes/No*)
- Relationship between HDL cholesterol and a continuous measure of waist circumference (*cm*)

Scatter plot of smoking vs. HDL



Stata command: `twoway (scatter hdl smoking)`

Scatter plot of waist circumference vs. HDL



Stata command: `twoway (scatter hdl waist)`

Linear regression

- Mean of outcome assumed to change linearly with a continuous predictor
- Uses all the data to estimate mean at any point; *borrows strength across points by assuming smoothness*
- Predictor can be of <u>any</u> type: *binary, categorical, count,* or *continuous*
- Distributional assumptions *not* required for predictor(s)
- Failure of linearity assumption can be addressed via transformations

Systematic part of linear model

- Model represents the mean of y as a function of x

$$\mathrm{E}\left[y|x\right] = \beta_0 + \beta_1 x$$

- Values of $\mathrm{E}[y \mid x]$ lie along the regression line
  - estimated values are known as the *fitted values*

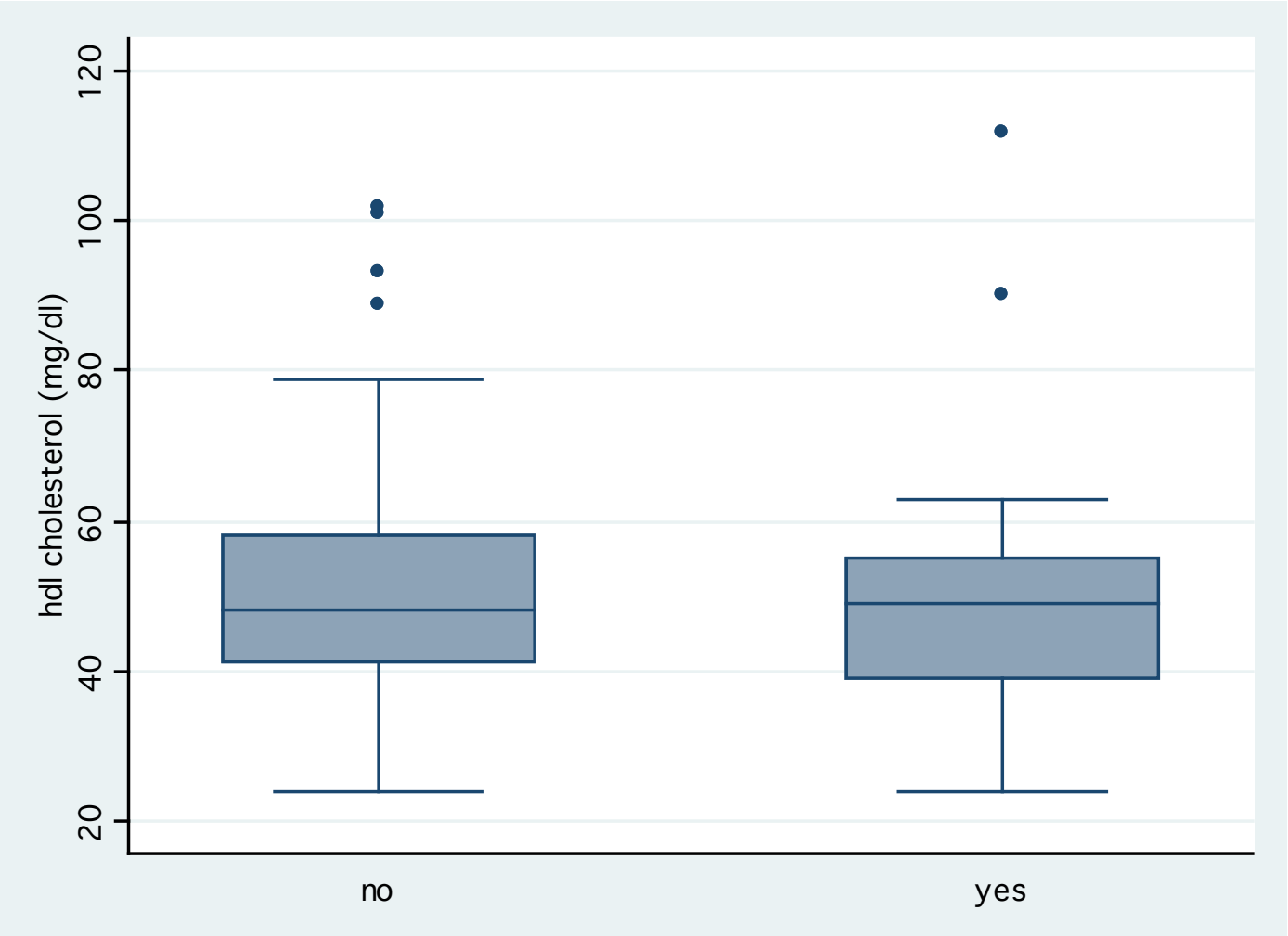
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- Dependence on predictor summarized by the slope of the line, β_1
  - β_1 measures change in outcome for a unit increase in x
- Intercept β_0 gives mean of y when $x = 0$

Random part of linear model

- $y = \beta_0 + \beta_1 x + \varepsilon$
  - observed outcome = conditional mean + random error
  - estimated values of ε are known as *residuals*
- Assumptions about random errors (ε)
  - independent
  - mean zero at all values of x
  - equal variance at all values of x
  - normally distributed (at least in small samples)
- These assumptions are required for valid inference (i.e. for *p*-values & confidence intervals)
- No assumptions about the predictor are required

HDL by smoking status



Stata command: `graph box hdl, over(smoking)`

(A more informative summary than the scatter plot on slide 23)

A formal comparison of the mean HDL levels between smoking groups can be made using the two-sample *t*-test:

```
. ttest hdl, by(smoking)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      no |     193    50.26943    .9490756    13.18498    48.39748    52.14138
     yes |      26    50.34615    3.578114    18.24487    42.97689    57.71542
---------+--------------------------------------------------------------------
combined |     219    50.27854    .9345828    13.83056    48.43656    52.12051
---------+--------------------------------------------------------------------
    diff |           -.0767238    2.895971               -5.784556    5.631109
------------------------------------------------------------------------------
    diff = mean(no) - mean(yes)                                   t =  -0.0265
Ho: diff = 0                     degrees of freedom =      217

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.4894         Pr(|T| > |t|) = 0.9789          Pr(T > t) = 0.5106
```

- Regression isn't needed for simple two-group comparisons of means

- A regression model for the above comparison yields identical results:

An simple linear regression model addressing the same hypothesis as the *t*-test from the HDL - smoking example:

```
. regress hdl i.smoking

      Source |       SS       df       MS              Number of obs =     219
-------------+------------------------------           F(  1,   217) =    0.00
       Model |   .13487973     1   .13487973            Prob > F      =  0.9789
    Residual |  41699.8743   217  192.165319            R-squared     =  0.0000
-------------+------------------------------           Adj R-squared = -0.0046
       Total |  41700.0091   218  191.284446            Root MSE      =  13.862

------------------------------------------------------------------------------
         hdl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   1.smoking |   .0767238   2.895971     0.03   0.979    -5.631109    5.784556
       _cons |   50.26943   .9978353    50.38   0.000     48.30274    52.23612
```

- Regression output provides the estimated diff. between groups, mean HDL in non-smoking group, but not mean HDL in smoking group. (Means for all groups can be obtained via *postestimation* commands in Stata.)

- Inference about the effect of smoking identical to the previous analysis

Interpretation of linear regression model for single binary predictor (smoking status):

Group	*Mean HDL*
Overall:	$E[y \mid x] = \beta_0+\beta_1 x$
smokers ($x=1$):	$\beta_0+\beta_1$
non-smokers ($x=0$):	β_0

Interpretation of β_1 : Difference in mean HDL between smokers and non-smokers

Note on categorical predictors in regression models:

- Binary predictors coded as $0/1$ can be entered directly in regression models without concerns about coding
- A linear regression model with a single binary predictor gives results equivalent to a two-sample *t*-test comparing the means in the groups specified by the predictor
- Categorical variables with >2 levels can be entered in a regression model via binary indicators for each level, reserving one as the reference category
- A linear regression for a single categorical variable with >2 levels is equivalent to a one-way analysis of variance

Postestimation commands in Stata

- After a regression model is fitted via `regress`, Stata saves a number of values related to the fit that can be used for *postestimation* analyses. These include:
  - Regression coefficients
  - Standard errors and model statistics
- Common postestimation analyses include
  - use of `predict` command to compute:
    - *fitted values*
    - *residuals*
    - *influence statistics*
  - use of `lincom` command to estimate means in specific subgroups defined by the model

Post-estimation commands in Stata

Example: using `predict` to estimate fitted values and residuals.

```
. predict yfit, xb

. predict yres, residuals

. list yfit yres hdl smoking if _n<=10
```

	yfit	yres	hdl	smoking
1.	50.26943	11.73057	62	no
2.	50.26943	-4.26943	46	no
3.	50.26943	-21.26943	29	no
4.	50.26943	1.73057	52	no
5.	50.34615	-1.346154	49	yes
6.	50.26943	-7.26943	43	no
7.	50.34615	12.65385	63	yes
8.	50.26943	-5.26943	45	no
9.	50.26943	-10.26943	40	no
10.	50.26943	2.73057	53	no

Note : model only yields two unique fitted values since `smoking` takes on only two values

Post-estimation commands in Stata

Example: using `lincom` to estimate the mean outcome value for smokers and non-smokers in the previous example (recall slide 17).

Issued immediately after a `regress` command, `lincom` provides estimates of `lin`ear `com`binations of the estimated regression coefficients, and also provides a std. error and confidence interval:

```
. lincom _cons + smoking
------------------------------------------------------------------------------
     smoking |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   50.34615   2.718635    18.52   0.000     44.98784    55.70446
------------------------------------------------------------------------------
. lincom _cons
------------------------------------------------------------------------------
         hdl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   50.26943   .9978353    50.38   0.000     48.30274    52.23612
------------------------------------------------------------------------------
```

Note : the second command returns the intercept

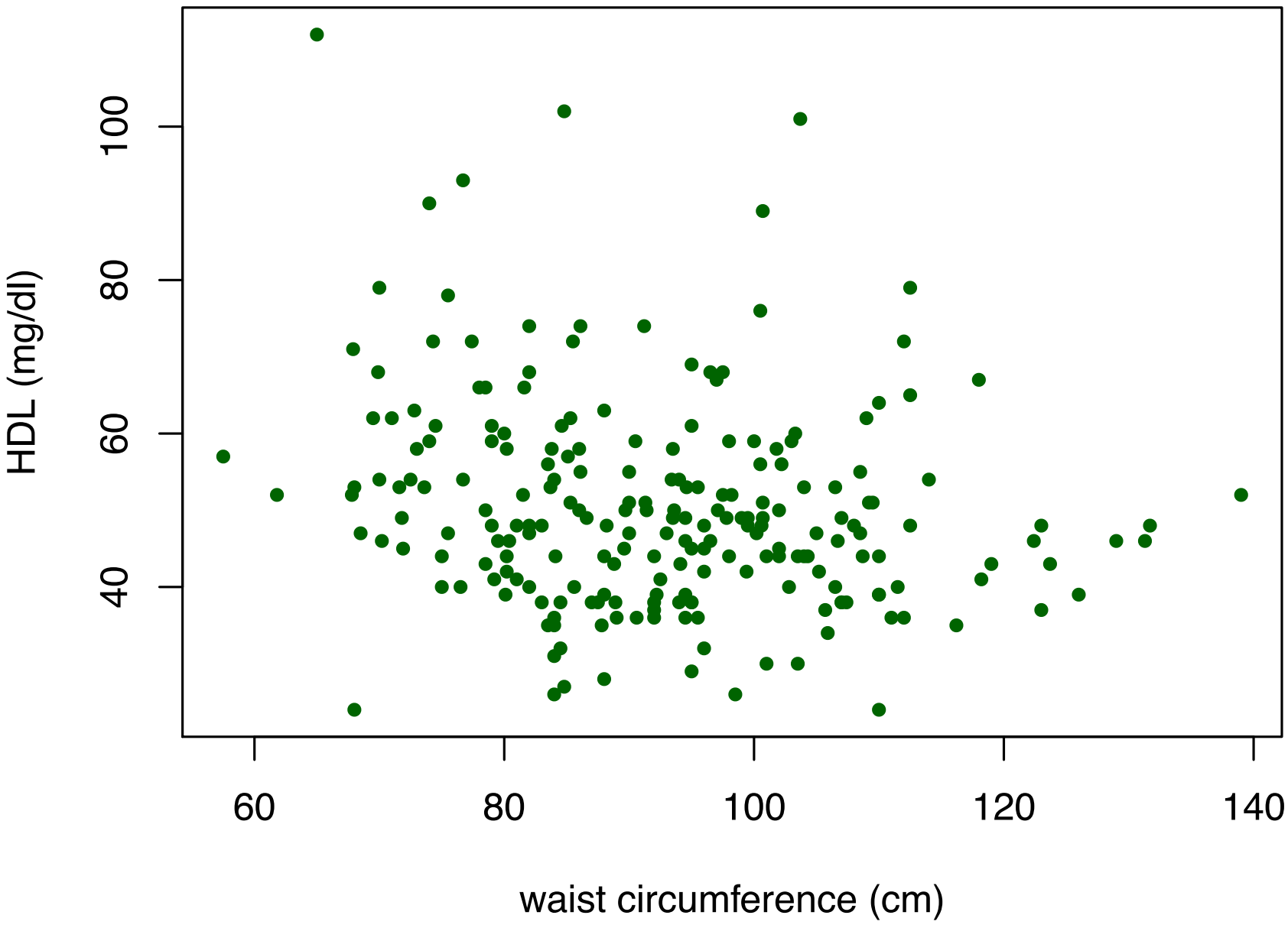
Simple linear regression with a continuous predictor

- A tool for understanding dependence of a the mean of a continuous outcome on a single continuous predictor
- With a continuous predictor, summary of dependence is a "line of means"

Example analysis topic:

Relationship between HDL (mg/dl) and the continuous measure waist circumference (cm)

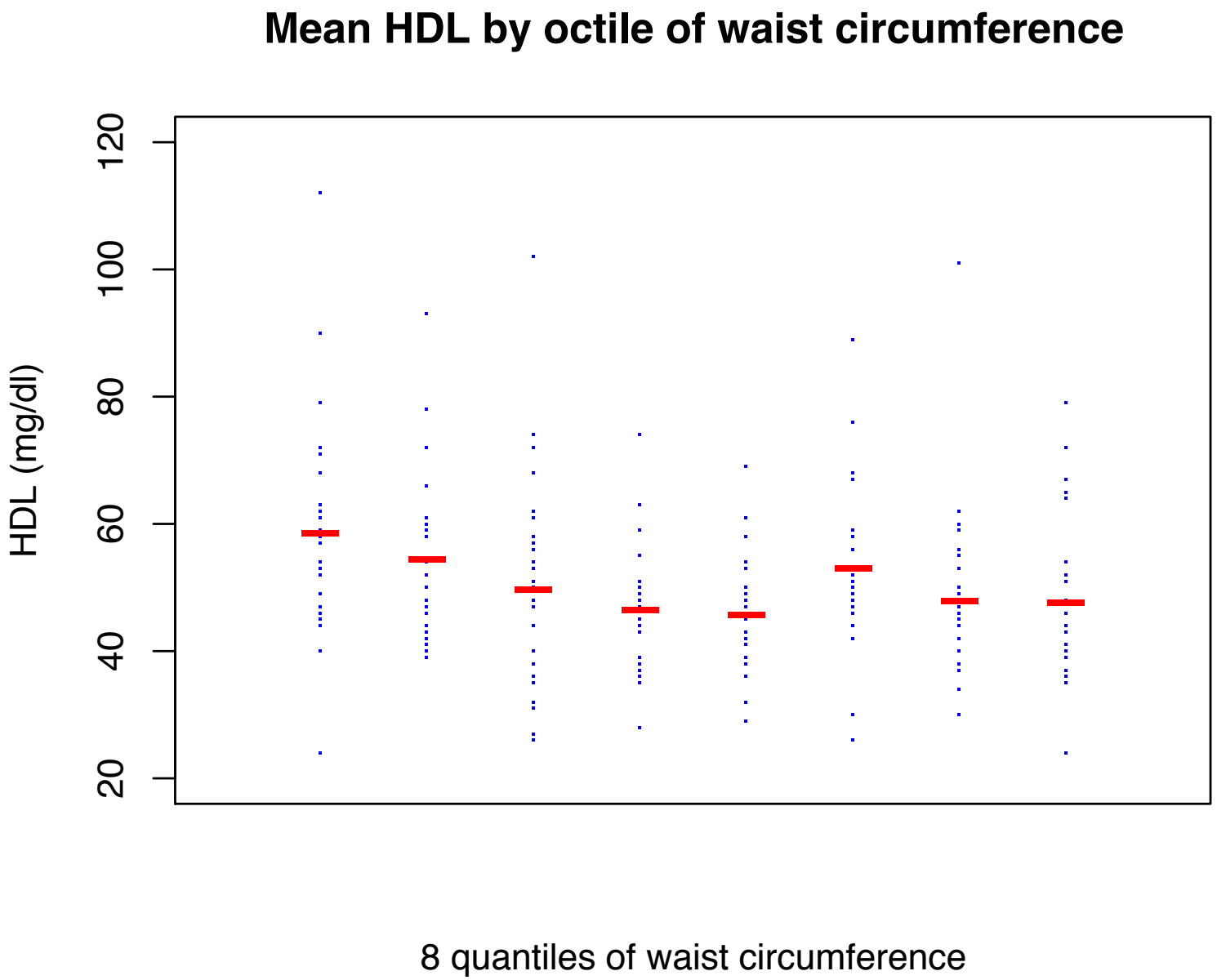
Scatter plot of waist circumference vs. HDL



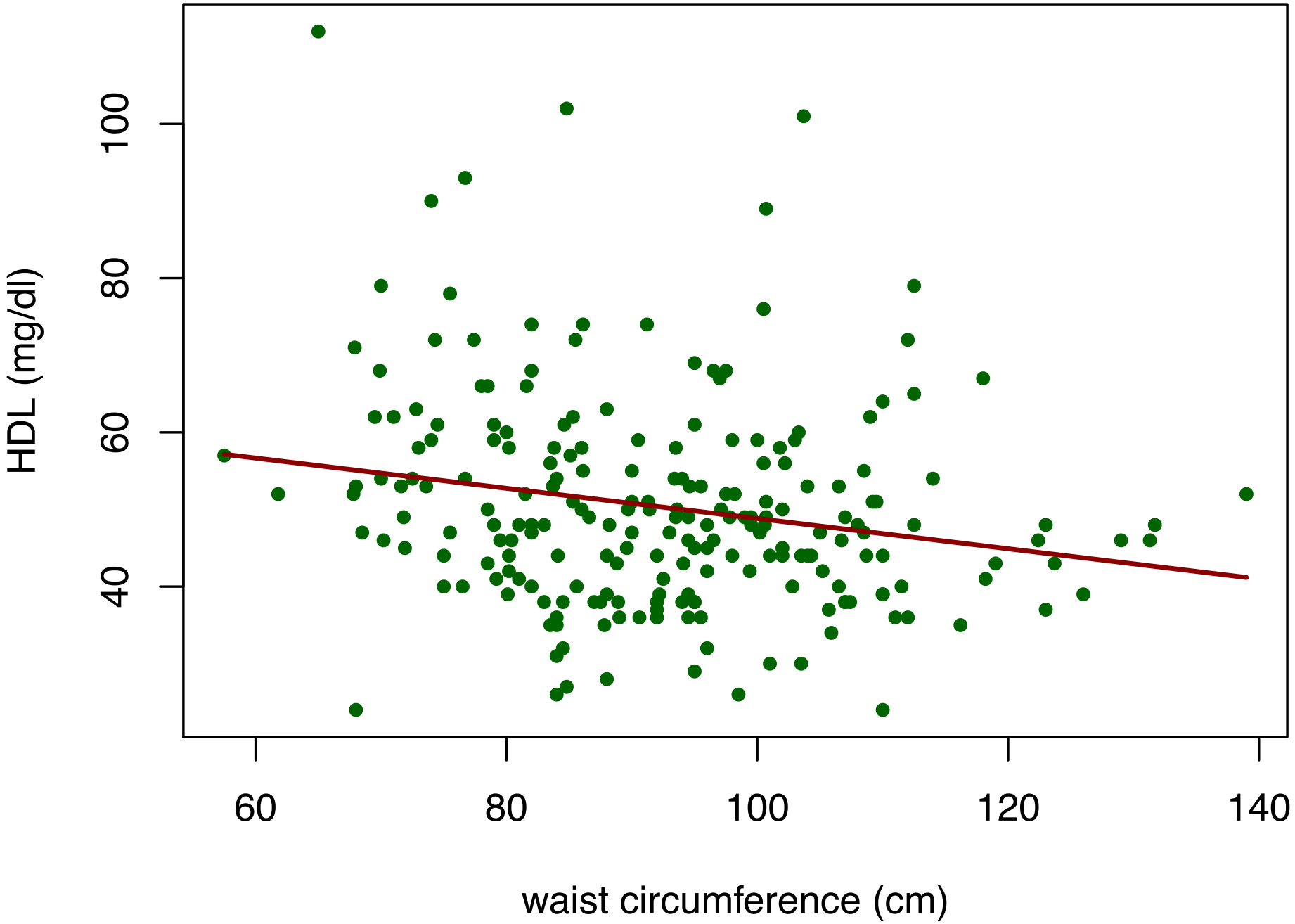
Stata command: `twoway (scatter hdl waist)`

Visualization of relationship without assuming a model:

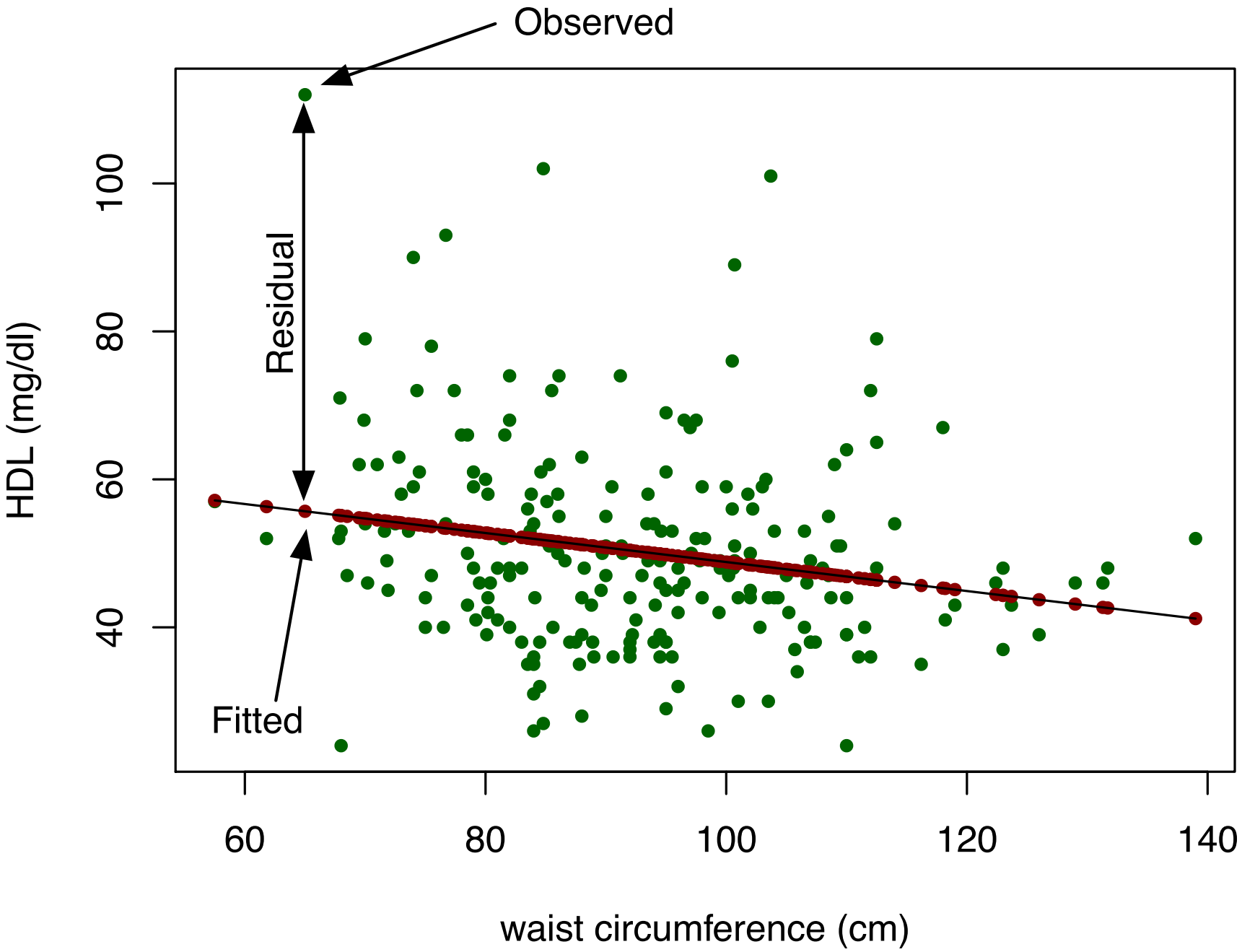
- plotting means in groups defined by quantiles of `waist`



Linear regression model: mean HDL is linearly related to waist circumference



Observed data, fitted value, residual



Ordinary least squares (OLS) estimation

- Slope and intercept estimated using OLS
- OLS estimates
  - determine the regression line
  *- minimize sum of squared residuals (*RSS*)*

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad \mathrm{MSS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad \mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Efficient - *minimum variance in class of estimators*
- Maximum likelihood estimates (MLEs) if outcome is normal
- Drawback of OLS: sensitive to outliers

Dependence of HDL on waist circumference

```
. regress hdl waist

      Source |       SS       df       MS              Number of obs =     219
-------------+------------------------------           F(  1,   217) =    9.66
       Model |  1777.68854     1  1777.68854            Prob > F      =  0.0021
    Residual |  39922.3206   217  183.973828            R-squared     =  0.0426
-------------+------------------------------           Adj R-squared =  0.0382
       Total |  41700.0091   218  191.284446            Root MSE      =  13.564

------------------------------------------------------------------------------
         hdl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       waist |   -.1960017   .0630536    -3.11   0.002    -.3202776   -.0717258
       _cons |    68.42785   5.910124    11.58   0.000     56.77925    80.07644
------------------------------------------------------------------------------
```

The sample estimate of β_1 gives estimated change in mean HDL for each unit increase in waist circ. (in cm)

Mean HDL decreases by about -0.2 mg/dL for each centimeter increase in waist circumference

Precision and statistical significance of slope estimate

```
. regress hdl waist

      Source |       SS       df       MS              Number of obs =     219
-------------+------------------------------           F(  1,   217) =    9.66
       Model |  1777.68854     1  1777.68854            Prob > F      =  0.0021
    Residual |  39922.3206   217  183.973828            R-squared     =  0.0426
-------------+------------------------------           Adj R-squared =  0.0382
       Total |  41700.0091   218  191.284446            Root MSE      =  13.564

------------------------------------------------------------------------------
         hdl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       waist |  -.1960017   .0630536    -3.11   0.002    -.3202776   -.0717258
       _cons |   68.42785   5.910124    11.58   0.000     56.77925    80.07644
------------------------------------------------------------------------------
```

The standard error of the estimate for β_1

Model *t*-statistics are calculated as the ratio coeff. / std. error. The associated hypothesis tests evaluate the null hypoth. that the true value of the coefficients are zero.

The *F* statistic evaluates the significance of the improvement in the model including the predictor relative to one that excludes the predictor (i.e. that only estimates the overall mean HDL)

Proportion of variation explained

```
. regress hdl waist

      Source |       SS       df       MS              Number of obs =     219
-------------+------------------------------           F(  1,   217) =    9.66
       Model |  1777.68854     1  1777.68854            Prob > F      =  0.0021
    Residual |  39922.3206   217  183.973828           R-squared     =  0.0426
-------------+------------------------------           Adj R-squared =  0.0382
       Total |  41700.0091   218  191.284446           Root MSE      =  13.564

------------------------------------------------------------------------------
         hdl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       waist |  -.1960017   .0630536    -3.11   0.002    -.3202776   -.0717258
       _cons |   68.42785   5.910124    11.58   0.000     56.77925    80.07644
------------------------------------------------------------------------------
```

- R-squared is the proportion of total variance in the outcome explained by the regression model:

$$1777.68854/41700.0091 = 0.0426$$

- Adjusted R-squared accounts for the number of predictors in the model

Marginal and residual variance estimates

```
. regress hdl waist

      Source |       SS       df       MS              Number of obs =     219
-------------+------------------------------           F(  1,   217) =    9.66
       Model |  1777.68854     1  1777.68854            Prob > F      =  0.0021
    Residual |  39922.3206   217  183.973828            R-squared     =  0.0426
-------------+------------------------------           Adj R-squared =  0.0382
       Total |  41700.0091   218  191.284446            Root MSE      =  13.564

------------------------------------------------------------------------------
         hdl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       waist |  -.1960017   .0630536    -3.11   0.002    -.3202776   -.0717258
       _cons |   68.42785   5.910124    11.58   0.000     56.77925    80.07644
------------------------------------------------------------------------------
```

- The total mean squared error (MS) is the variance of the outcome

- Root mean squared error (MSE) is an index of model accuracy

Interpreting coefficients from the linear model

- Model represents the mean of HDL as a function of waist circumference
  - mean(*HDL* | *waist circ.*) = 68.43 - 0.196 X (*waist circ.)*
- 68.43 gives mean HDL for an individual with waist circumference of 0 cm
- -0.196 gives change in aver. HDL assoc. with a 1 cm increase in waist circ.
  - mean(*HDL* | *waist circ.* = 100) - mean(*HDL* | *waist circ.* = 99) = [68.43 - 0.196 X (100)] - [68.43 - 0.196 X (99)] = -0.196
- Use model to compute change in aver. HDL assoc. with a 10 unit increase in waist circ.
  - mean(*HDL* | *waist circ.* = 100) - mean(*HDL* | *waist circ.* = 90) = [68.43 - 0.196 X (100)] - [68.43 - 0.196 X (90)] = -0.196 X 10 = -1.96

Regression postestimation commands

- Issued after fitting a regression model
- Used for model assessment and interpretation

Example: using `lincom` to estimate the mean outcome for an individual with waist size of 120 cm.

```
. lincom _cons + waist*120

 ( 1)  120*waist + _cons = 0

------------------------------------------------------------------------------
         hdl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   44.90764   1.955863    22.96   0.000     41.05272    48.76257
------------------------------------------------------------------------------
```

Example: using `lincom` to estimate the change in mean outcome for a 10 cm increase in waist size.

```
. lincom  waist*10

 ( 1)  10*waist = 0

------------------------------------------------------------------------------
         hdl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -1.960017   .6305362    -3.11   0.002    -3.202776   -.7172577
------------------------------------------------------------------------------
```

Postestimation commands

Example: The `predict` command is used to estimate fitted values and residuals.

```
. predict yfit, xb
. predict yres, residuals
(2 missing values generated)

. list if _n<=10
     +-----------------------------------------------+
     | smoking   waist   hdl       yfit        yres |
     |-----------------------------------------------|
  1. |      no    85.3    62    51.7089     10.2911 |
  2. |      no   106.7    46   47.51447   -1.514468 |
  3. |      no      95    29   49.80769   -20.80769 |
  4. |      no    98.2    52   49.18048    2.819518 |
  5. |     yes   100.7    49   48.69048    .3095219 |
     |-----------------------------------------------|
  6. |      no    88.8    43    51.0229   -8.022897 |
  7. |     yes      88    63    51.1797     11.8203 |
  8. |      no     102    45   48.43568   -3.435675 |
  9. |      no    85.6    40    51.6501    -11.6501 |
 10. |      no    71.6    53   54.39413   -1.394127 |
     +-----------------------------------------------+
```

Note : model yields unique fitted values for distinct values of `smoking` and `waist`

Questions a simple linear regression can answer

1. How does mean of outcome depend on predictor?
2. What is the precision of the estimated association, and is it statistically significant?
3. How much of overall variation in the outcome is explained by the estimated linear model?

Summarizing regression results

"In unadjusted analysis, higher values of waist circumference were associated with lower average levels of HDL (0.2 mg/dL lower HDL per cm increase in waist circumference; 95% CI -0.32 to -0.07, p = 0.002). However, waist circumference accounted for only 4% of the variability in HDL."

A good summary ....

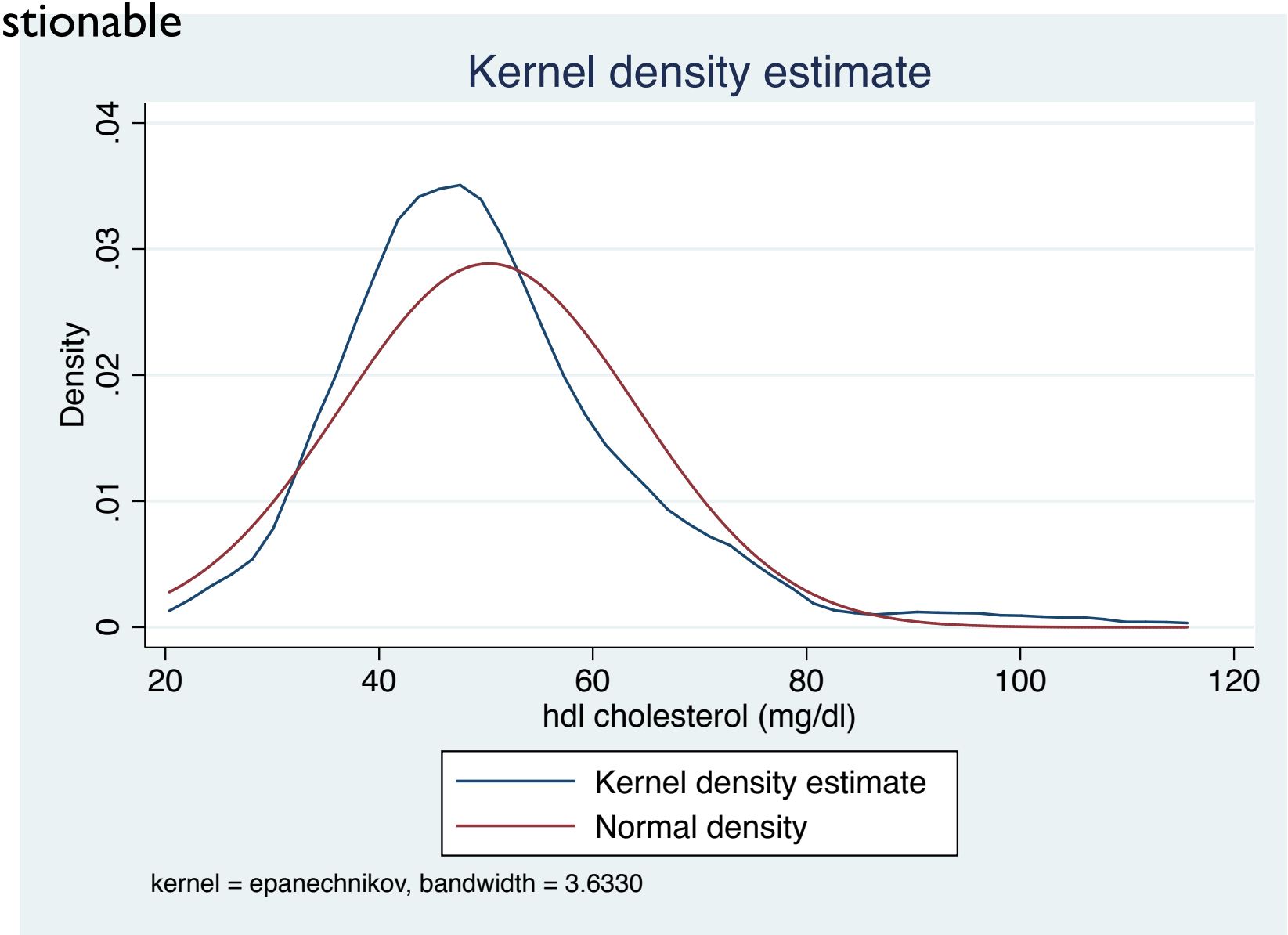
- Doesn't just focus on statistical significance
- Cites slope estimate *and* 95% CI
- Uses sensible "units" for slope
- May use *R*-squared as a measure of substantive importance

Special Topic: Bootstrap confidence intervals

- When assumptions required for statistical inference from regression models are violated, the *bootstrap* can be used to obtain valid inferences (Section 3.6, VGSM)
- Bootstrapping involves drawing repeated samples with replacement from the original sample, and estimating the quantities of interest from each. The variability in these estimates is used to obtain confidence intervals
- Bootstrap confidence intervals are "nonparametric" in the sense that they avoid distributional assumptions typically required for inferences from regression models
- Typically, a large number of samples (e.g. 1000) is required

Example: Bootstrap confidence interval for HDL - waist circumference model

- Exploratory analysis reveals that normality of the distribution of HDL is questionable



Stata command: `kdensity hdl, normal`

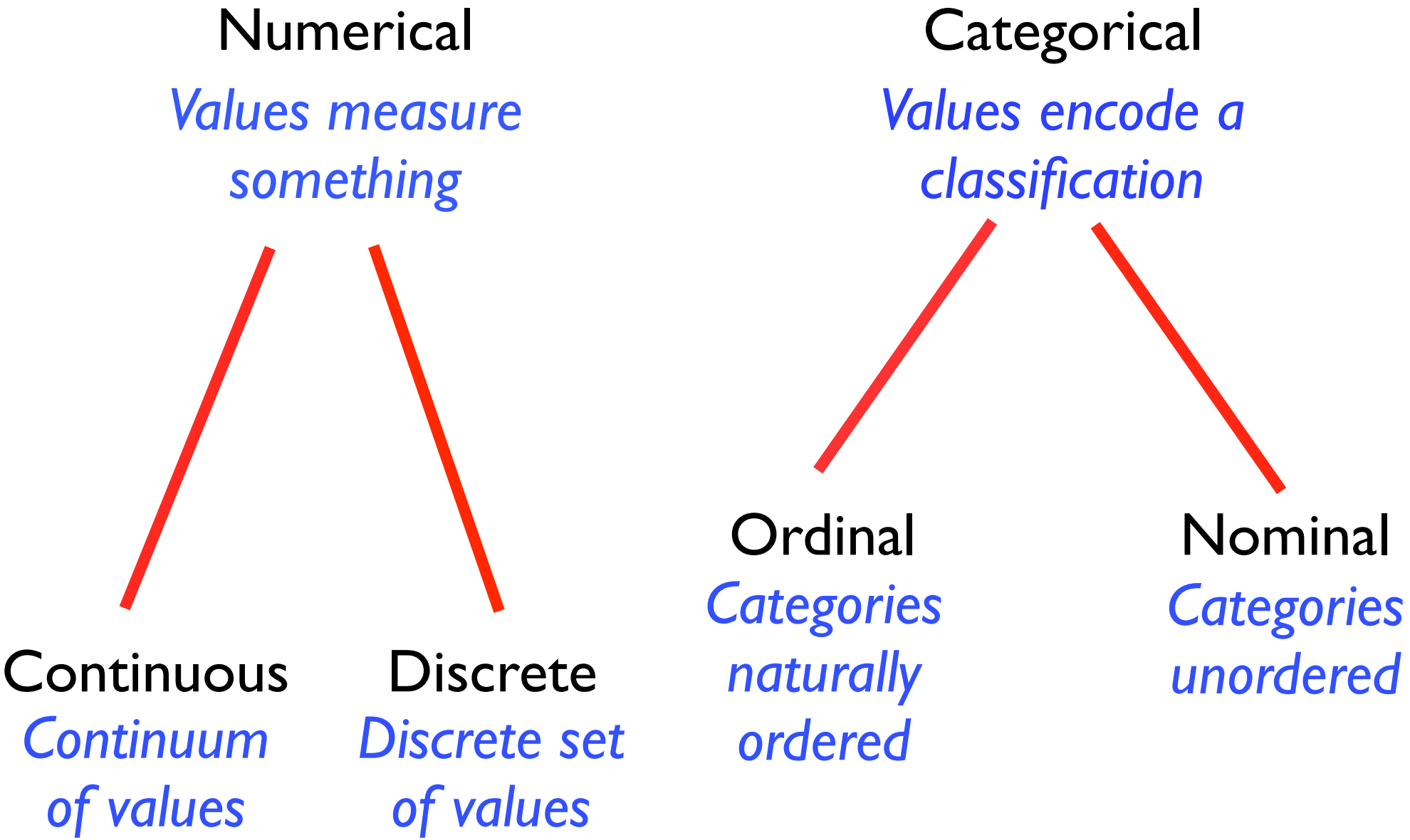
Example: Bootstrap confidence interval for HDL - waist circumference model

```
. bootstrap `"regress hdl waist"' _b, reps(1000)

command:      regress hdl waist
statistics:   b_waist    = _b[waist]
              b_cons     = _b[_cons]

Bootstrap statistics                                  Number of obs    =       219
                                                      Replications     =      1000

------------------------------------------------------------------------------
Variable     |  Reps  Observed      Bias  Std. Err. [95% Conf. Interval]
-------------+----------------------------------------------------------------
     b_waist |  1000 -.1960017 -.0012388  .0667359  -.3269603  -.0650432   (N)
             |                                     -.3373865  -.0717307   (P)
             |                                     -.3401462  -.0727492   (BC)
      b_cons |  1000  68.42785  .1313524   6.41104    55.8472    81.0085   (N)
             |                                      56.19201   81.75299   (P)
             |                                      56.11214   81.73902   (BC)
------------------------------------------------------------------------------
Note:  N   = normal
       P   = percentile
       BC  = bias-corrected
```

- In this case, inferences largely unchanged compared to standard approach (see slide #43)

Brief Review of Exploratory Data Analysis

Data types


Numerical
Values measure something
Categorical
Values encode a classification
Continuous
Continuum of values
Discrete
Discrete set of values
Ordinal
Categories naturally ordered
Nominal
Categories unordered

Uses of data exploration

- Find data errors
- Assess missingness
- Detect anomalous observations and outlying data values
- Select appropriate analysis methods
- Support a formal data analysis

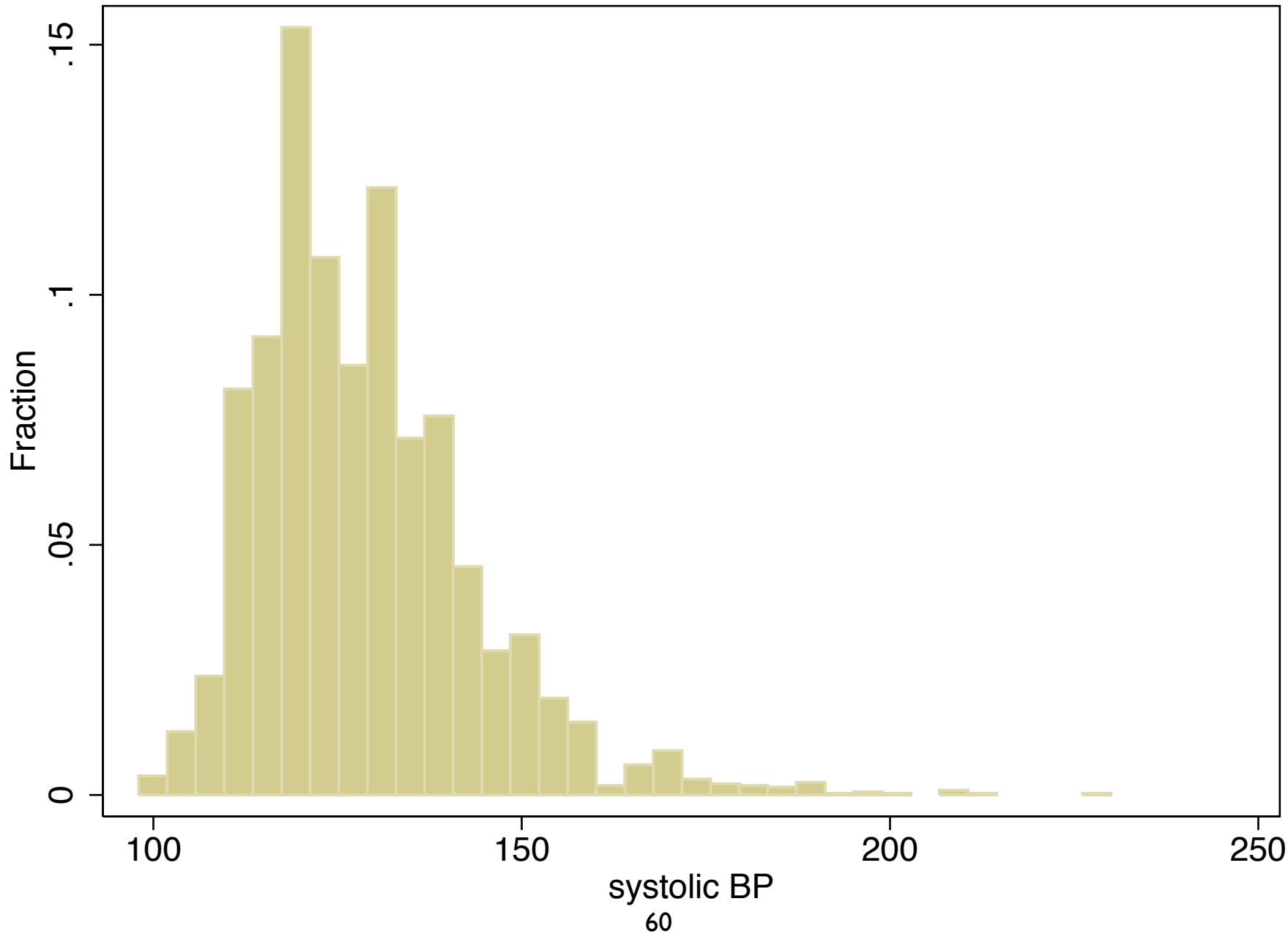
Graphical data summaries

- Can effectively communicate the distribution
- Methods have different strengths:
  - Histogram: *captures location, spread, shape of the distribution; "density" plot as a smoothed histogram*
  - Boxplot: *gives 5-number summary, shows outliers, skewness*
  - Normal Q-Q plot: *assesses Normality*

Histogram

- Shows location, spread, and shape of the distribution
- Horizontal axis: *intervals or "bins" in which data values are grouped*
- Vertical axis: *number, fraction, or percent of the observations in each bin*

Histogram of SBP



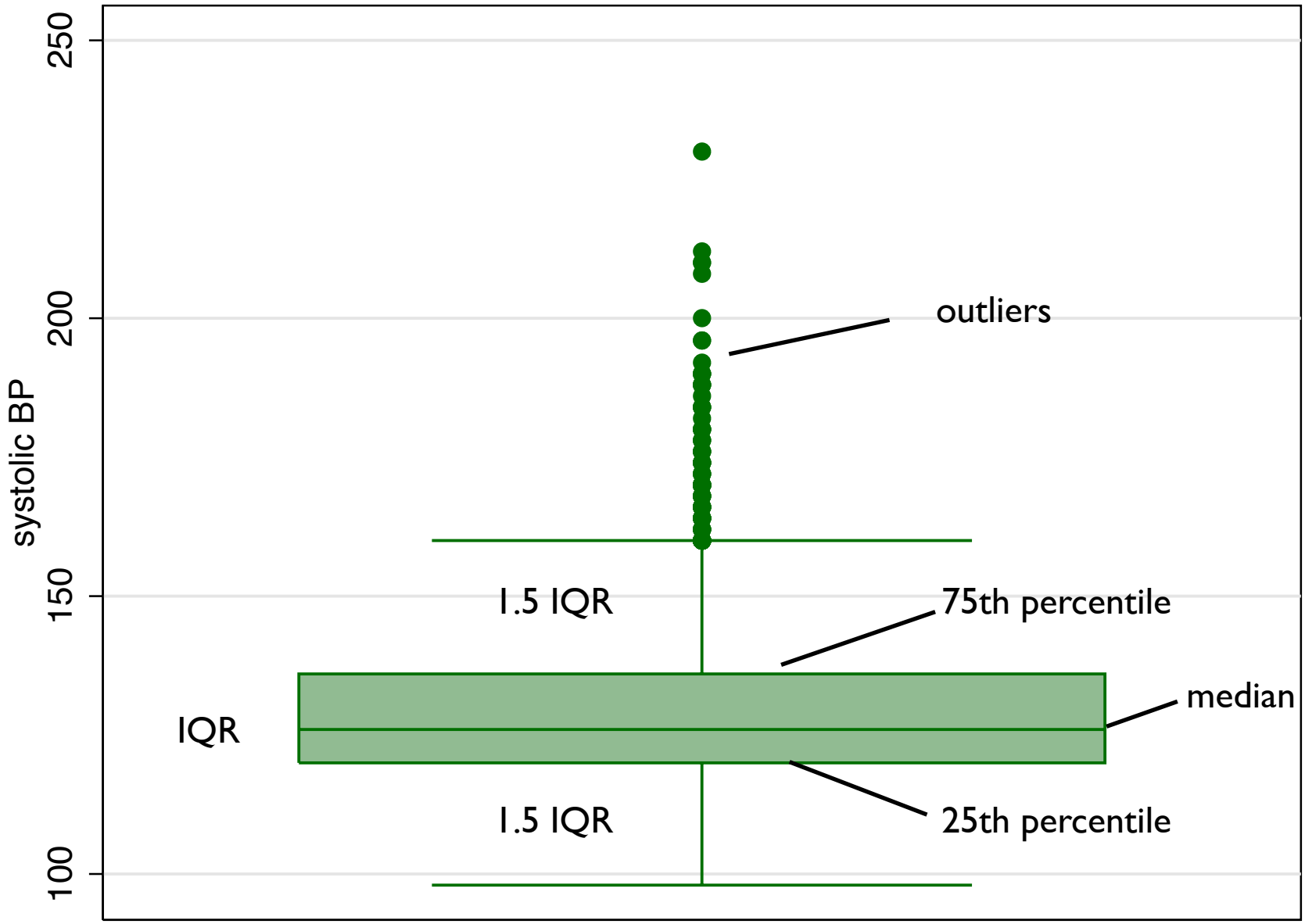
Interpreting Histograms

- Pattern of bar heights conveys shape of distribution:
  - *number of modes*
  - *skewness*
  - *long or short tails*
- Usefulness depends on number of bins
  - *too many defeats goal of summarization*
  - *too few obscures shape of distribution*

Boxplots

- Upper and lower limits of box: 25th and 75th percentiles (upper and lower quartiles); *height of box is interquartile range (IQR)*
- Median drawn as a line across the box
- 'Whiskers' at 1.5 IQRs beyond the quartiles (or at maximum/minimum)

Boxplot of Systolic Blood Pressure

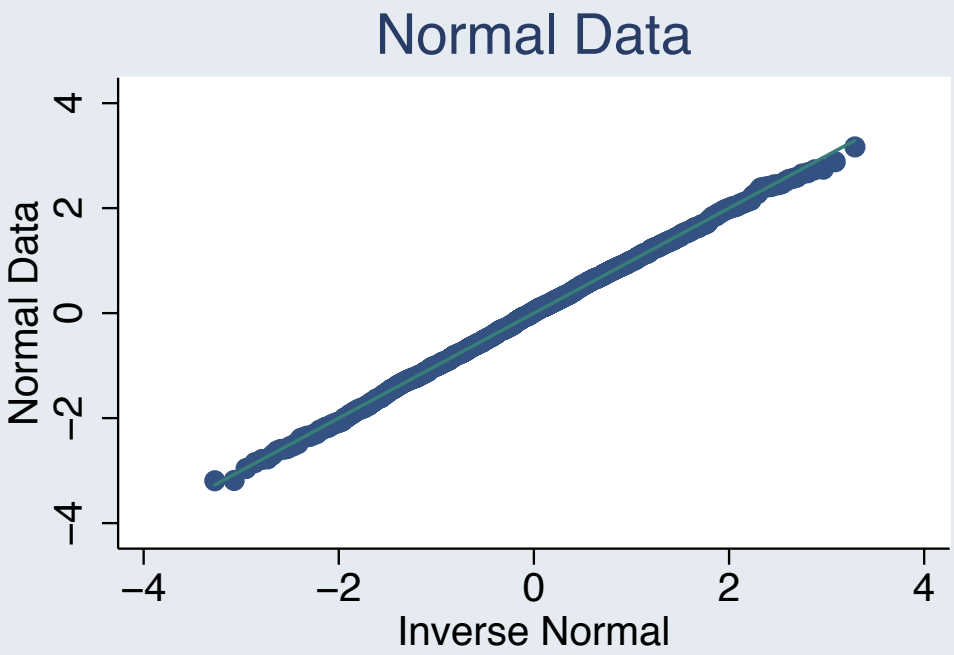


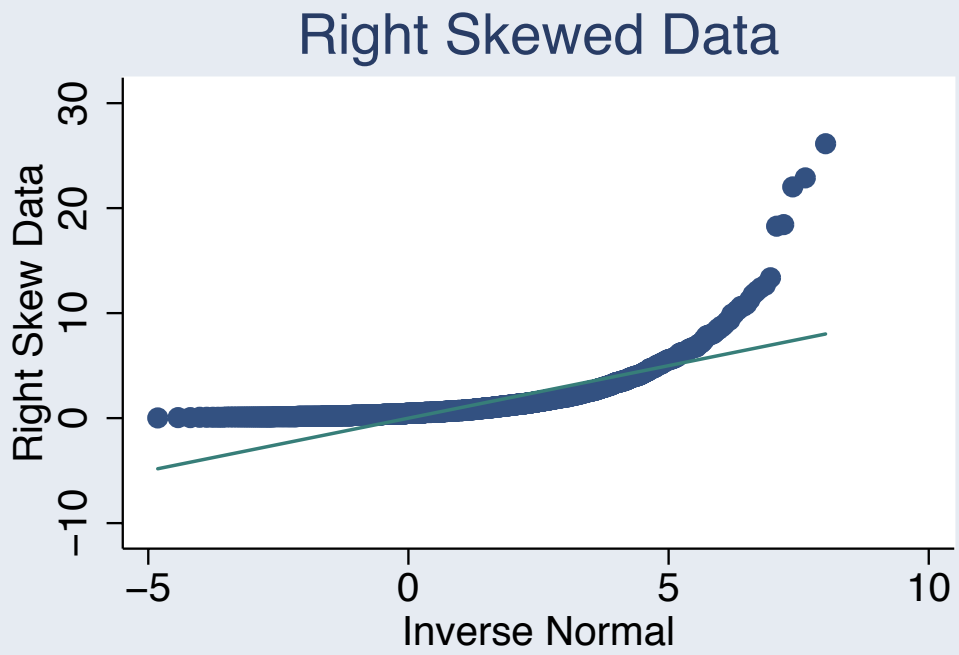
Interpreting a boxplot

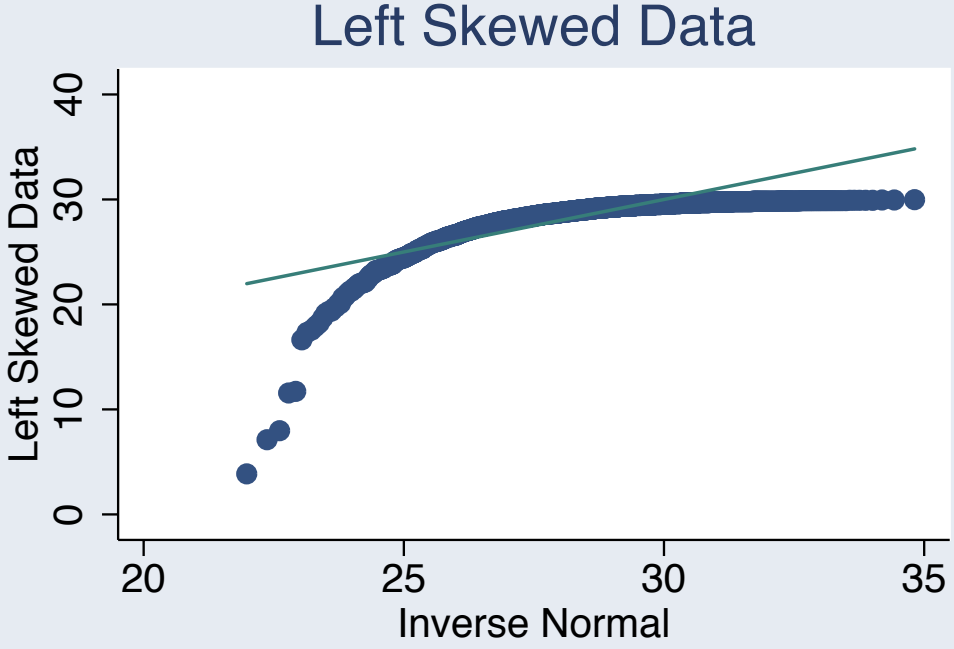
- Location: the *median*
- Spread: height of box (*IQR); also by extent of outliers*
- Skewness: *is the median halfway between the quartiles? Is one of the whiskers shorter than the other? Are most of the outliers on one side?*
- Outliers: values beyond whiskers

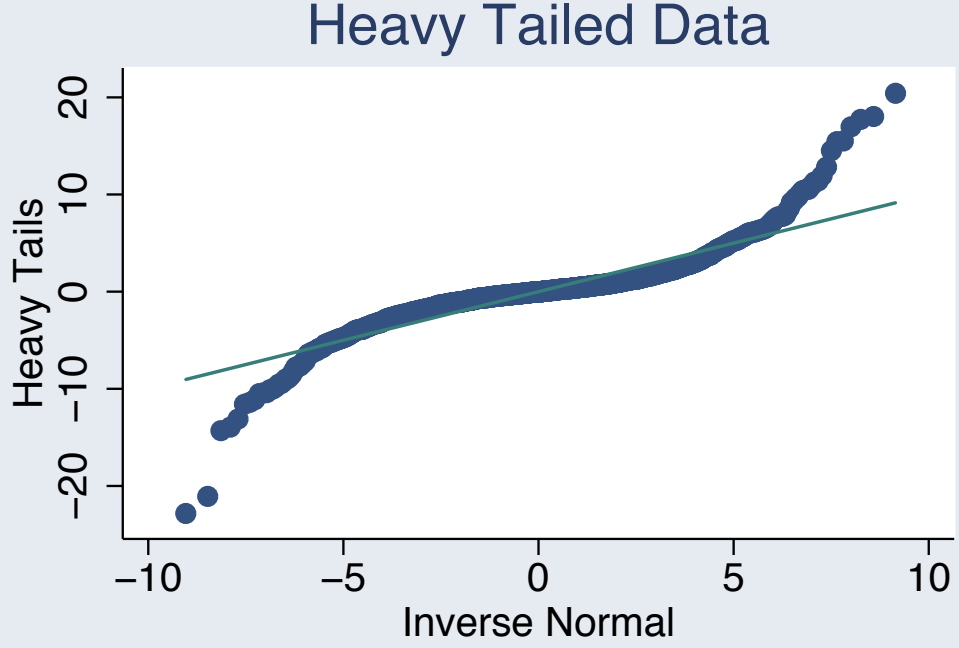
Normal Q-Q Plot

- Graphical assessment of Normality
- Vertical axis: *actual data values*
- Horizontal axis: *corresponding "quantiles" from a normal distribution with same mean, and SD*
- If points lie along a roughly straight line, data are approximately normal
  - *usually noisy in tails, especially with small N*
- Curvature indicates nature of departure from normality


Normal Data
Normal Data
−4
−2
0
2
4
−4
−2
0
2
4
Inverse Normal


Right Skewed Data
Right Skew Data
−10
0
10
20
30
−5
0
5
10
Inverse Normal


Left Skewed Data
Left Skewed Data
0
10
20
30
40
20
25
30
35
Inverse Normal


Heavy Tailed Data
Heavy Tails
−20
−10
0
10
20
−10
−5
0
5
10
Inverse Normal

Interpreting Q-Q Plots

- Right skewness: plot curved up
- Left skewness: plot curved down
- Light or heavy tails: S-curvature
- Outliers: values far above line on right, below it on the left
- STATA command for Q-Q plot for SBP: `qnorm sbp`