Linear Regression

January 15, 2019
Biostatistics 208

This week:

- Review of simple linear regression model
- Centering & scaling predictors for interpretability
- Outliers and assessing linearity in the simple linear model
- Introduction to multiple linear regression
- Post-estimation in multiple predictor models

Simple linear regression model

- Model represents the conditional mean of y as a function of x
  - $E(y \mid x) = \beta_0 + \beta_1 x$
- Individual outcomes:
  - $y = \beta_0 + \beta_1 x + \varepsilon$
  - Observed outcome = conditional mean + random error
- Assumptions about random errors (ε):
  - independent
  - mean zero at all values of x
  - equal (i.e. constant) variance at all values of x
  - normally distributed (at least in small samples)

Review: The normality assumption

- Validity of inferences for linear regression requires that residuals (ε) be normally distributed with zero mean and constant variance (σ^2)

$$\varepsilon = y - (\beta_0 + \beta_1 x)$$

  - implies that y follows a normal distribution *conditional* on the predictor values (for fixed values of x, y differs from ε by a constant)

    Example - For binary x, two normal distributions are specified:

    For $x = 0$, $y = \varepsilon + \beta_0 \implies y \sim$ normal(mean= β_0, var=σ^2)

    For $x = 1$, $y = \varepsilon + \beta_0 + \beta_1 \implies y \sim$ normal(mean= $\beta_0 + \beta_1$, var=σ^2)

  - because of the *central limit theorem*, validity is of decreasing concern with increasing sample sizes (i.e. the distribution of coefficients tends to a normal distribution with increasing n)

  - however, the *constant variance assumption is required* even for large samples

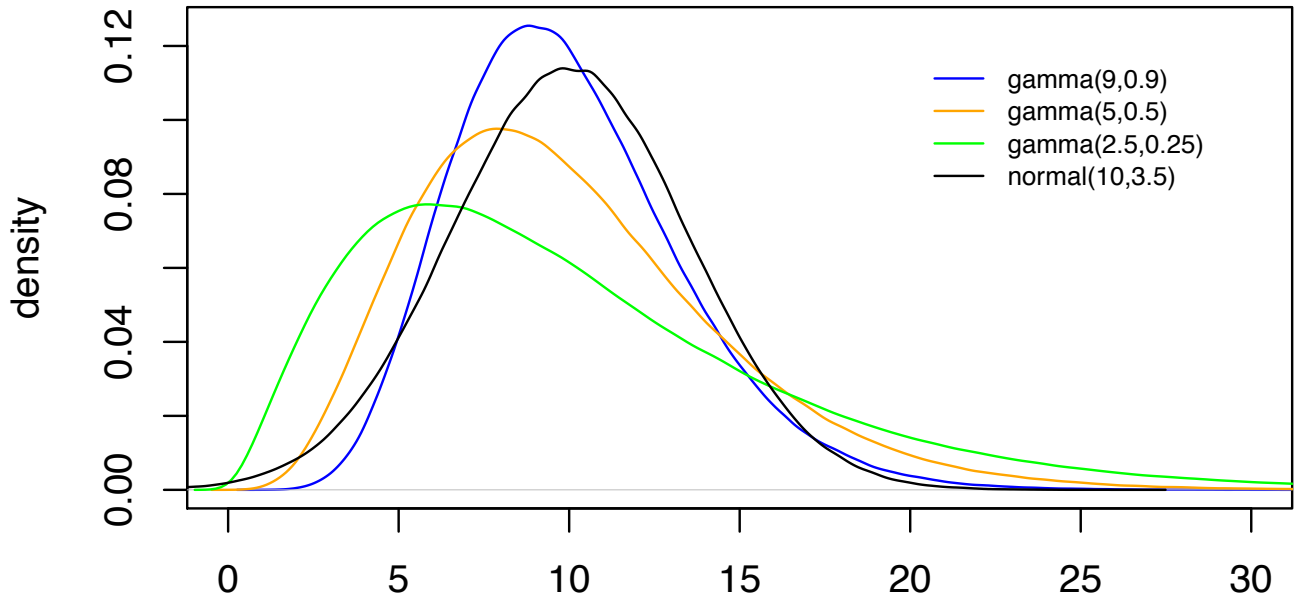
Review: Tests for non-normality

- There are a number of statistical tests for non-normality of a variable (e.g. Shapiro-Wilk, Kolmogorov-Smirnov)
  - null hypothesis: source distribution of the variable is the normal
- Some issues with the testing approach:
  - reject for small deviations from normality in large samples
    - in large samples, deviations from normality are typically less of an issue, so why bother testing?
  - reduced power for small samples

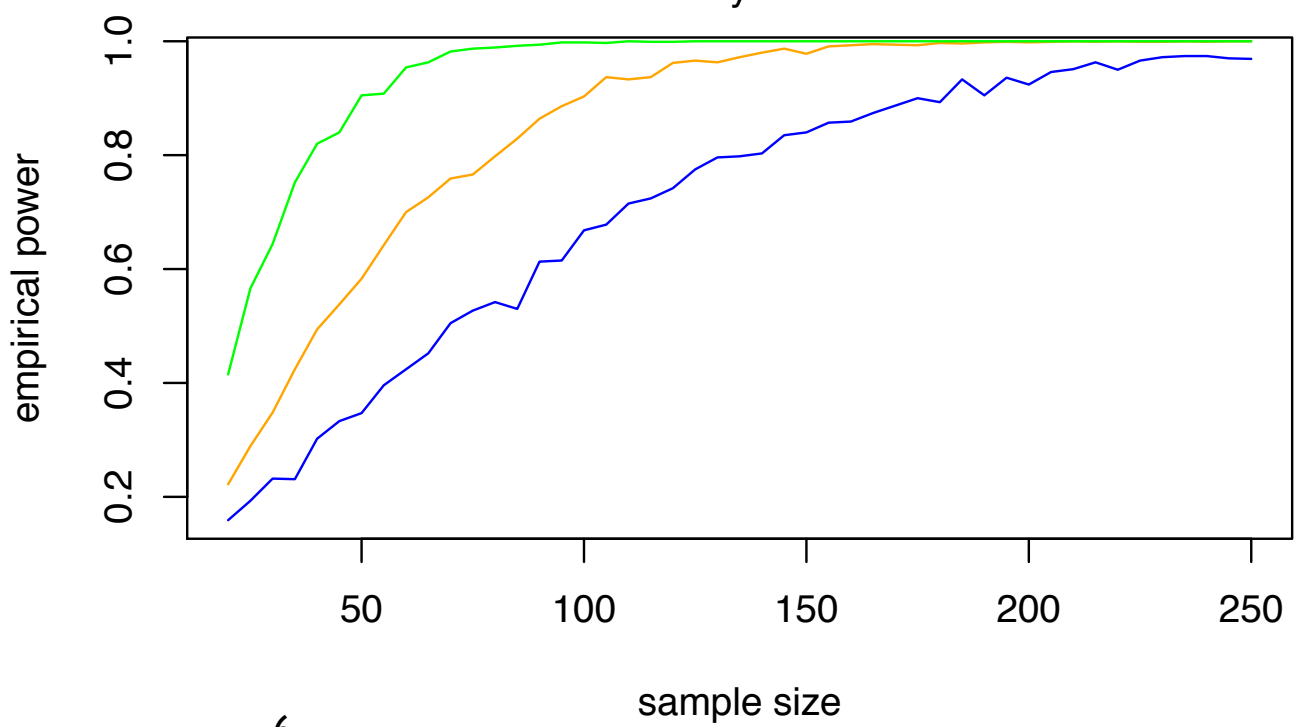
Reference: Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health. 2002;23:151-69.

Tests for non-normality - power

- Example: assume data (y) are generated from a range of non-normal (gamma) distributions with different skewness:



- Simulated power of *Shapiro-Wilk* test as a function of sample size:



Tests for non-normality

- Recommendations (see section 4.7.2.2 of text):
  - testing may be useful *in addition to* graphical assessment
  - not useful for large samples ($n \geq 100$)
  - if there's any question, consider an alternate means of inference (e.g. Wilcoxon (rank-based) test, bootstrap CI), at least as a sensitivity analysis

Note: in linear regression models the residuals (ϵ), not the outcome are assumed to be (approximately) normally distributed. In this context graphical assessment of the outcome is still sensible (section 4.7.2.3 of text) in suggesting transformations and identifying outliers

Example: Dependence of HDL on waist circumference in HERS (full sample at baseline)

```
. regress HDL waist

      Source |       SS       df       MS              Number of obs =    2750
-------------+------------------------------           F(  1,  2748) =  168.58
       Model |  27749.7008     1  27749.7008            Prob > F      =  0.0000
    Residual |  452355.921  2748  164.612781            R-squared     =  0.0578
-------------+------------------------------           Adj R-squared =  0.0575
       Total |  480105.622  2749  174.647371            Root MSE      =   12.83

------------------------------------------------------------------------------
         HDL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       waist |  -.2340937   .0180299   -12.98   0.000    -.2694471   -.1987403
       _cons |   71.73885   1.671793    42.91   0.000     68.46075    75.01694
------------------------------------------------------------------------------
```

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 -0.23 mg/dL for each centimeter increase in waist circumference (95% CI: -0.27, -0.20)

```
. regress HDL waist

      Source |       SS       df       MS              Number of obs =    2750
-------------+------------------------------           F(  1,  2748) =  168.58
       Model |  27749.7008     1  27749.7008            Prob > F      =  0.0000
    Residual |  452355.921  2748  164.612781            R-squared     =  0.0578
-------------+------------------------------           Adj R-squared =  0.0575
       Total |  480105.622  2749  174.647371            Root MSE      =   12.83

------------------------------------------------------------------------------
         HDL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       waist |  -.2340937   .0180299   -12.98   0.000    -.2694471   -.1987403
       _cons |   71.73885   1.671793    42.91   0.000     68.46075    75.01694
------------------------------------------------------------------------------
```

71.74 mg/dL is the estimated mean HDL for an indiv. with zero waist circ. (in cm)

Centering to make intercept meaningful

- In HDL example, intercept is meaningless
  - *no one has waist circumference of zero*
  - *represents an extrapolation beyond the data*
- Center predictor on the sample mean
  -*intercept gives mean HDL for waist circumference at sample mean*
- Slope estimate unaffected

Model with centered waist circumference

```
. egen meanwaist = mean(waist) if HDL!=.
(11 missing values generated)

. gen cwaist = waist - meanwaist
(13 missing values generated)

. reg HDL cwaist

      Source |       SS       df       MS              Number of obs =    2750
-------------+------------------------------           F(  1,  2748) =  168.58
       Model |  27749.7008     1  27749.7008            Prob > F      =  0.0000
    Residual |  452355.921  2748  164.612781            R-squared     =  0.0578
-------------+------------------------------           Adj R-squared =  0.0575
       Total |  480105.622  2749  174.647371            Root MSE      =   12.83

------------------------------------------------------------------------------
         HDL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      cwaist |  -.2340937   .0180299   -12.98   0.000    -.2694471   -.1987403
       _cons |   50.26655   .2446614   205.45   0.000     49.78681    50.74628
------------------------------------------------------------------------------
```

Exclude missing values on outcome

- estimated mean HDL for an indiv. with average waist circumference is 50.27 mg/dL (95% CI: 49.8, 50.7)

- note that 13 observations with missing HDL / waist are dropped

Scaling to make slope interpretable

- In HDL example, slope for waist circumference gives change in HDL for each *one* cm increase in waist size
- Creating a scaled predictor for waist size allows estimation of the effect on the outcome of a larger increase
- Intercept estimate unaffected

Model scaled waist circumference

```
. gen waist_10 = waist/10

. reg HDL waist_10

      Source |       SS       df       MS              Number of obs =    2750
-------------+------------------------------           F(  1,  2748) =  168.58
       Model |  27749.7002     1  27749.7002           Prob > F      =  0.0000
    Residual |  452355.922  2748  164.612781           R-squared     =  0.0578
-------------+------------------------------           Adj R-squared =  0.0575
       Total |  480105.622  2749  174.647371           Root MSE      =   12.83

------------------------------------------------------------------------------
         HDL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    waist_10 |  -2.340937   .1802986   -12.98   0.000    -2.694471   -1.987402
       _cons |   71.73884   1.671793    42.91   0.000     68.46075    75.01694
------------------------------------------------------------------------------
```

- HDL decreases by ~2.3 mg/dl for every 10 cm increase in waist size (95% CI: -2.7, -2.0)

Centering & Scaling

- Centering *and* scaling predictors affects both intercept and slope coefficients

Model with centered and scaled waist circumference

```
. gen cswaist = (waist - meanwaist)/10

. reg HDL cswaist

      Source |       SS       df       MS              Number of obs =    2750
-------------+------------------------------           F(  1,  2748) =  168.58
       Model |  27749.7008     1  27749.7008            Prob > F      =  0.0000
    Residual |  452355.921  2748  164.612781            R-squared     =  0.0578
-------------+------------------------------           Adj R-squared =  0.0575
       Total |  480105.622  2749  174.647371            Root MSE      =   12.83

------------------------------------------------------------------------------
         HDL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     cswaist |  -2.340937   .1802986   -12.98   0.000    -2.694471   -1.987403
       _cons |   50.26655   .2446614   205.45   0.000     49.78681    50.74628
------------------------------------------------------------------------------
```

- HDL decreases by ~2.3 mg/dl for every 10 cm increase in waist size
- Intercept gives estimated average HDL for an indiv. with average waist circumference

Model with standardized waist circumference

```
. egen stdwaist = std(waist) if HDL!=.

. reg HDL stdwaist

      Source |       SS       df       MS              Number of obs =    2750
-------------+------------------------------           F(  1,  2748) =  168.58
       Model |  27749.7008     1  27749.7008            Prob > F      =  0.0000
    Residual |  452355.921  2748  164.612781            R-squared     =  0.0578
-------------+------------------------------           Adj R-squared =  0.0575
       Total |  480105.622  2749  174.647371            Root MSE      =   12.83

------------------------------------------------------------------------------
         HDL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    stdwaist |   -3.17718   .2447059   -12.98   0.000    -3.657006   -2.697354
       _cons |   50.26655   .2446614   205.45   0.000     49.78681    50.74628
------------------------------------------------------------------------------
```

- The `std` option to `egen` generates standardized values of predictor variables (default is to center using the mean, and scale by the SD)
- -3.177 is the estimated change in mean HDL for each SD-unit increase in waist circ.

Scatterplots as model diagnostics

- Most useful when both outcome and predictor are continuous. Used to:
  - Detect outliers
  - Check linearity assumption, determine need for transformation
  - Diagnostic methods will be covered in more detail in lecture 6

Outliers and influential points

- An *outlier* is an extreme data point that lies far from the main spread of points in a scatter plot.
- *Influential points* can have a large impact on the estimated slope coefficient in a regression model
- An outlier is likely to be 'influential' if:
  - predictor value is anomalous (*a high leverage point)*
  - residual is large: *observed outcome is far from predicted value on regression line*
  - dataset is relatively small

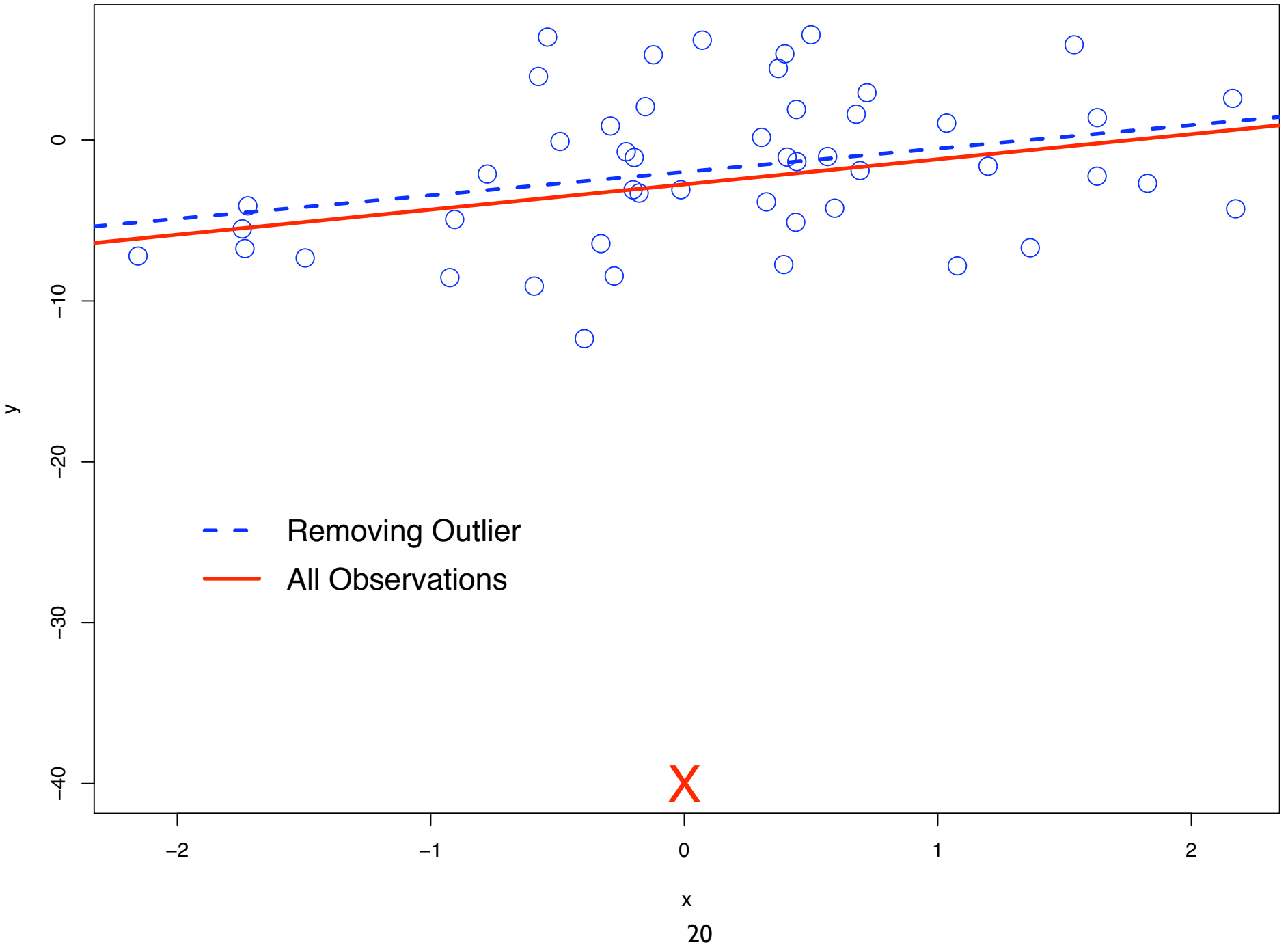
Three types of outliers

- Benign outlier:
  - *x not an outlier, y far from regression line*
  - *influences the intercept but not the slope*

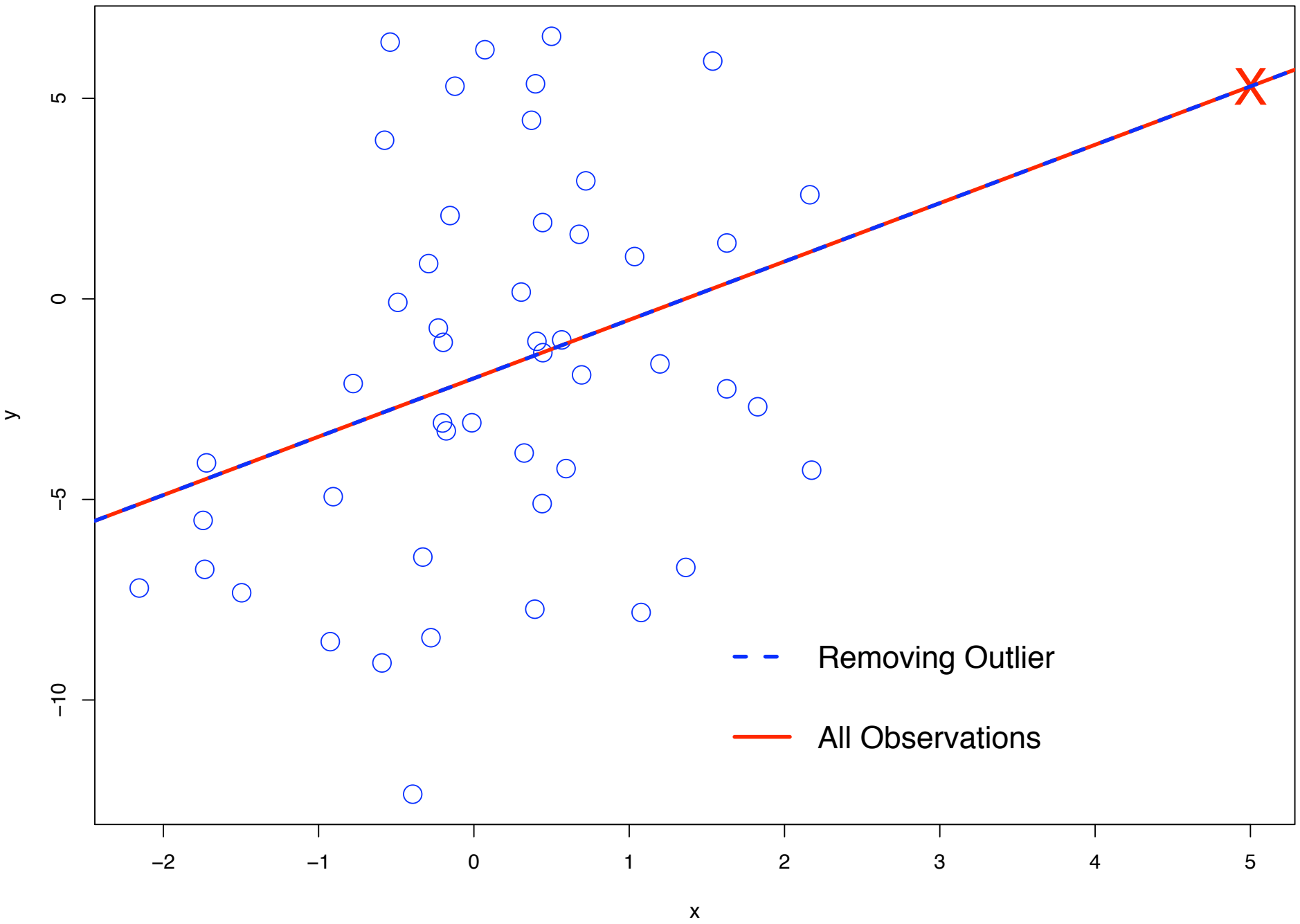
- High leverage point:
  - *x an outlier, y near regression line*
  - *influences neither slope nor intercept*
  - *may affect precision of estimated coefficients* and R^2

- Influential point:
  - *x an outlier, y far from regression line*
  - *strongly affects the slope (focus of interest)*
  - *this is the type of outlier to worry most about*

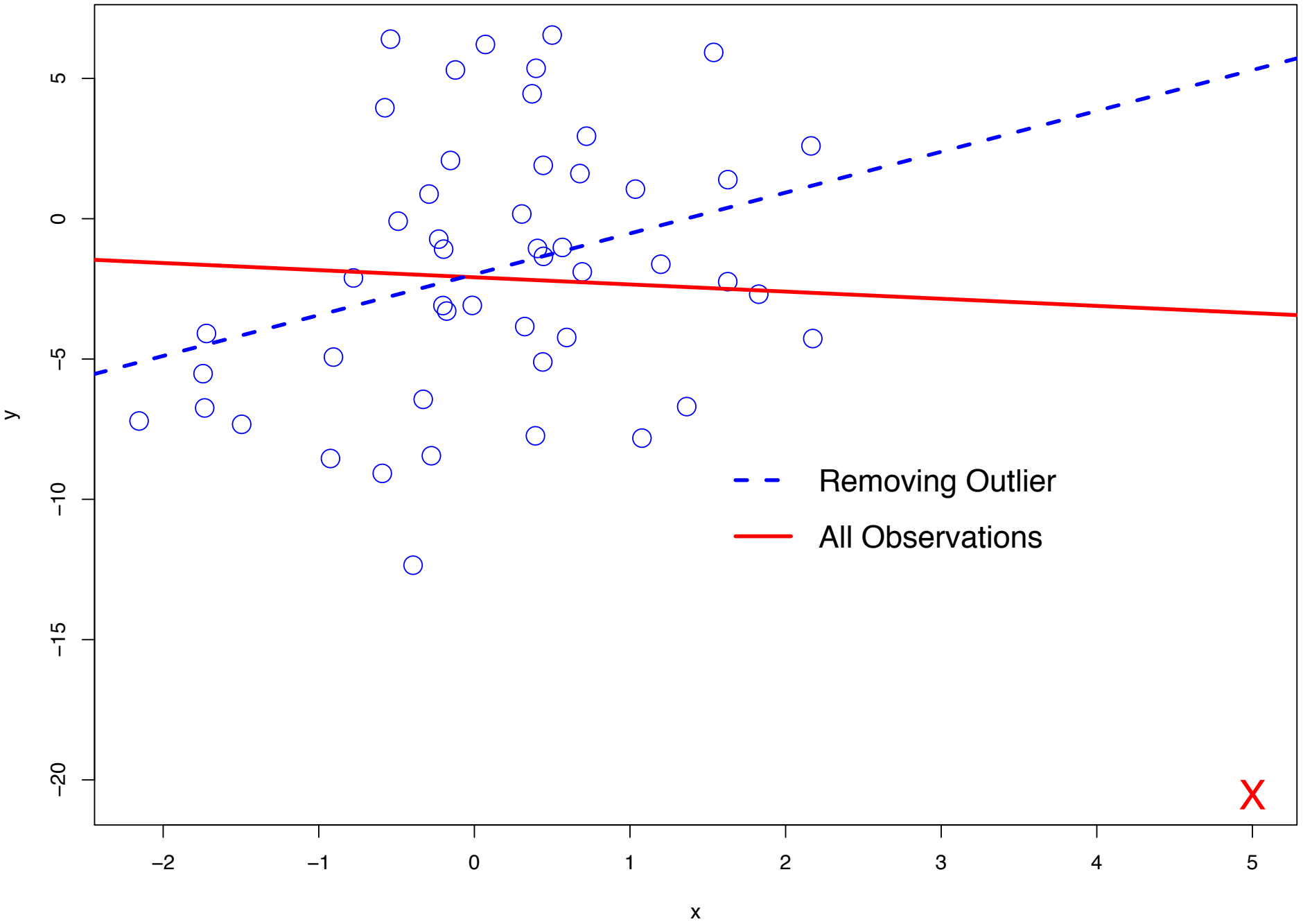
Example: Benign outlier



Example: High leverage point



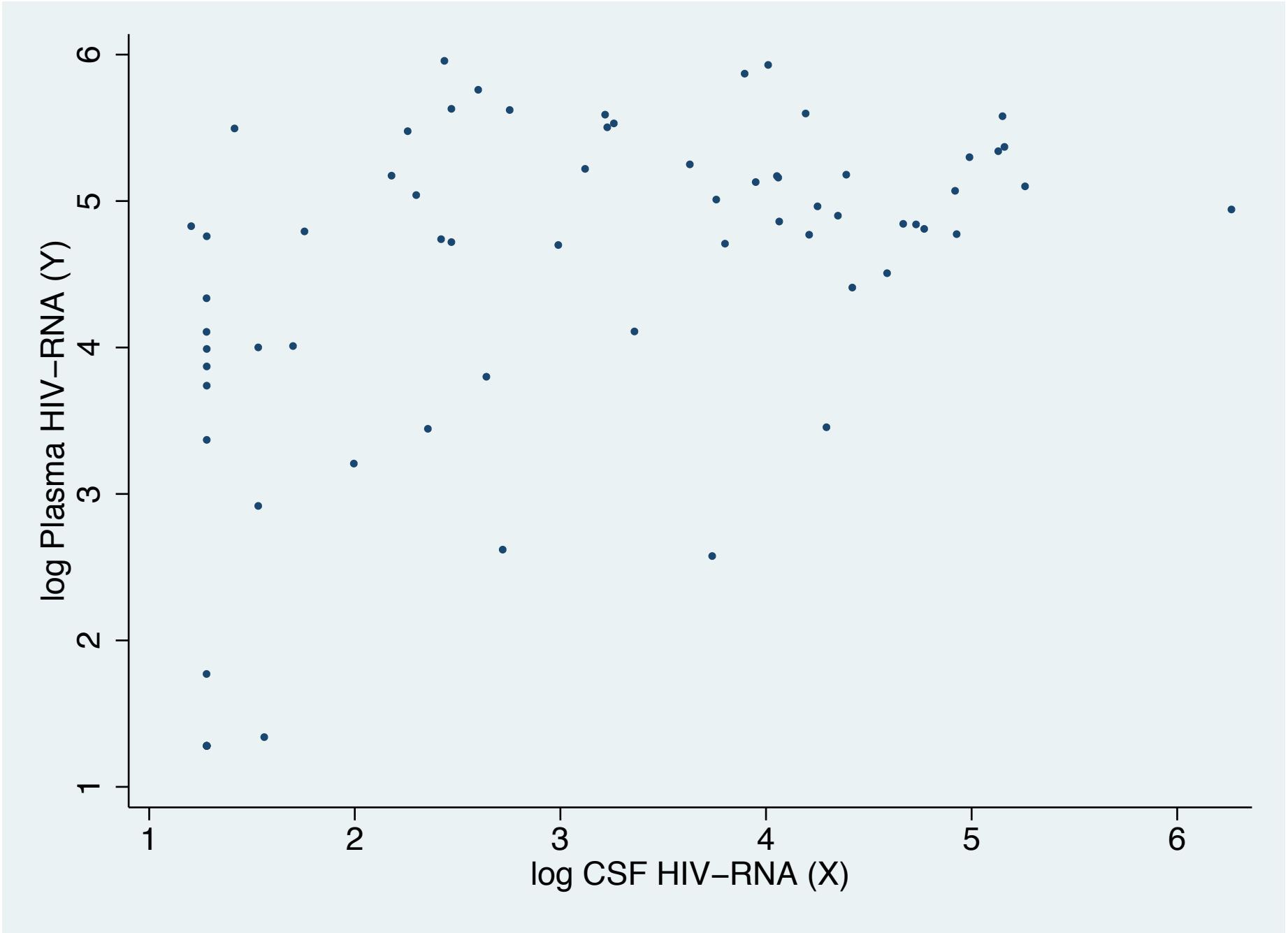
Example: Influential point



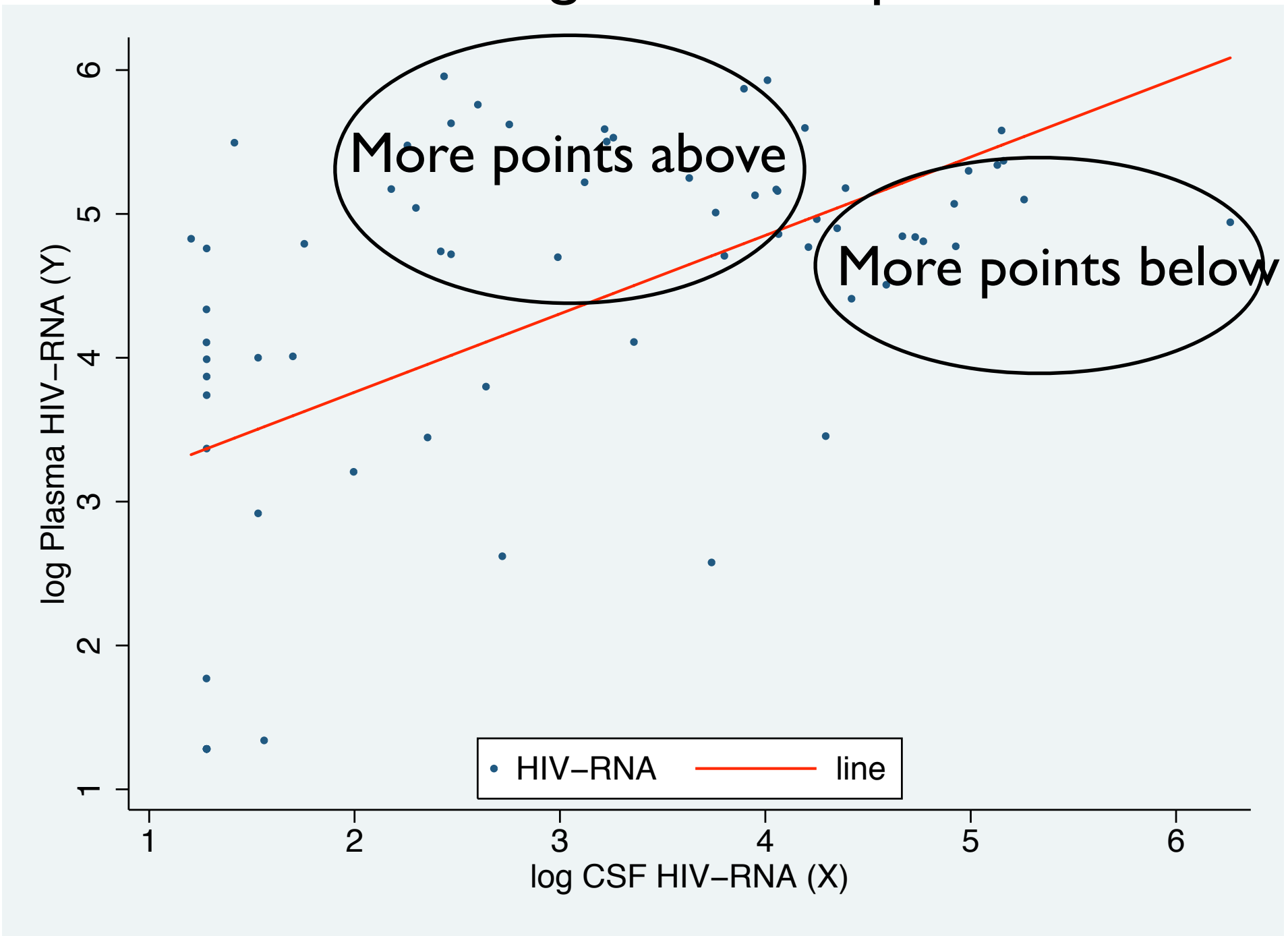
Exploratory methods for assessing linearity

- Linear regression with a continuous predictor assumes that the "line of means" is a straight line
- Scatterplots and nonparametric regression methods can help us diagnose violations of this assumption
- Nonlinear relationships can be often "linearized" using transformations of predictor(s) (lecture 6)
- Exploratory methods help visualize the relationship and suggest a course of action for modeling (more complicated for multiple predictors)

Example: relationship between HIV viral load in plasma and CSF



Is a straight line adequate?



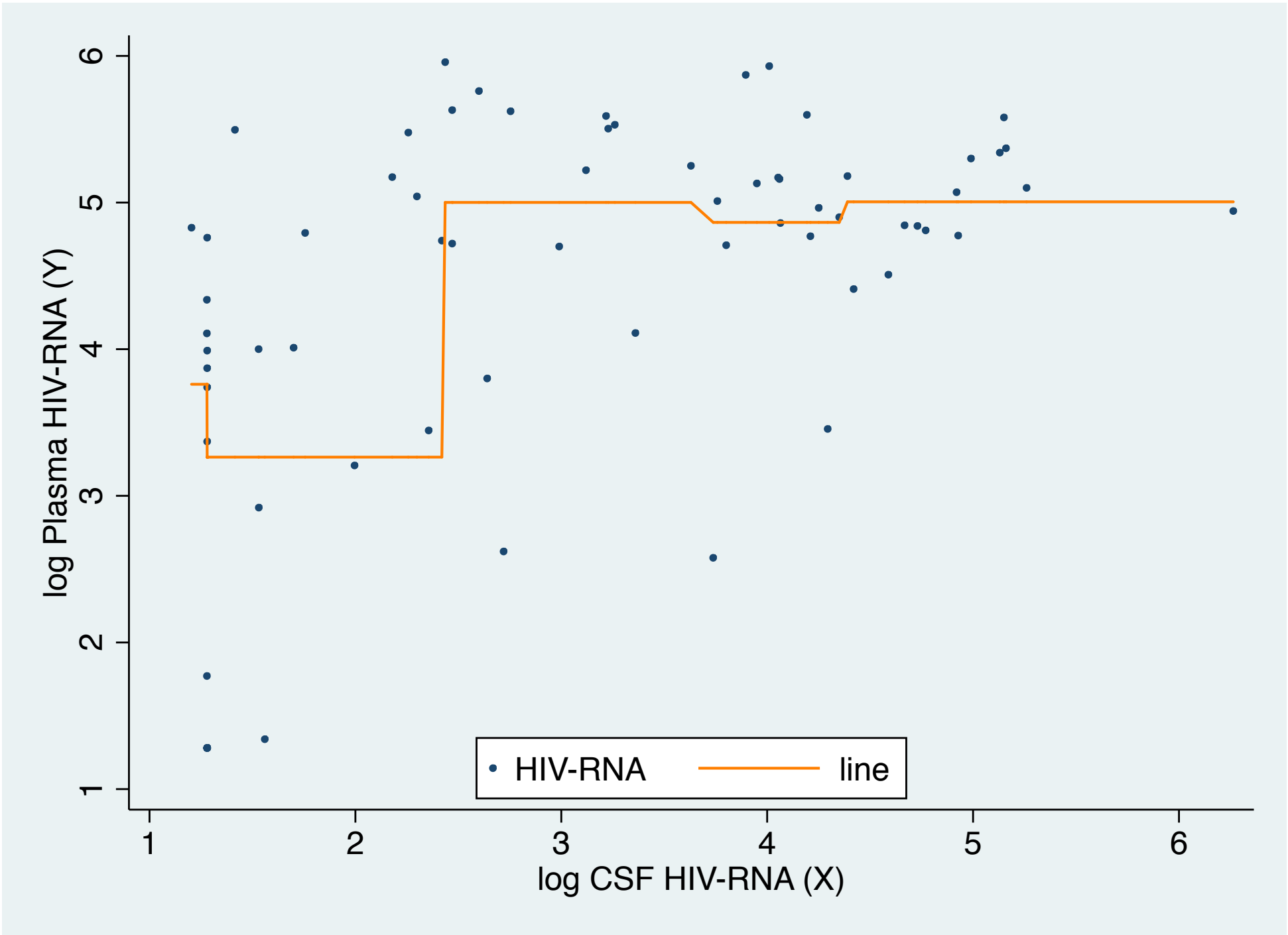
Step Function

- Split continuous predictor into several categories with similar numbers of observations in each
- Use the category-specific outcome means to represent the relationship between the average outcome and the predictor
- Result: a "step function"
- Estimates average value of outcome as a function of predictor, without assuming that the relationship is linear
- Depends on the categories chosen

Stata code to estimate and plot a step function approximation for the dependence of average log HIV viral load on CSF RNA level

```
* Generate a 5-category version of the predictor
. egen csfrna5 = cut(csfrna), group(5) label
* Fit a linear model (suppress output)
. quietly regress hivrna i.csfrna5
* Generate the predicted outcome using the fitted model
. predict hivrna2
(option xb assumed; fitted values)

* Plot result (next slide)
. twoway (scatter hivrna csfrna , msize(vsmall)) (line hivrna2
csfrna, sort clpat(solid) lcolor(orange)),
plotregion(style(none)) ytitle("log Plasma HIV-RNA (Y)")
xtitle("log CSF HIV-RNA (X)") legend(order(1 2) label(1 "HIV-
RNA") label(2 "line") pos(6) ring(0)) saving(csf3, replace)
```


log Plasma HIV-RNA (Y)
1
2
3
4
5
6
log CSF HIV-RNA (X)
1
2
3
4
5
6
HIV-RNA
line

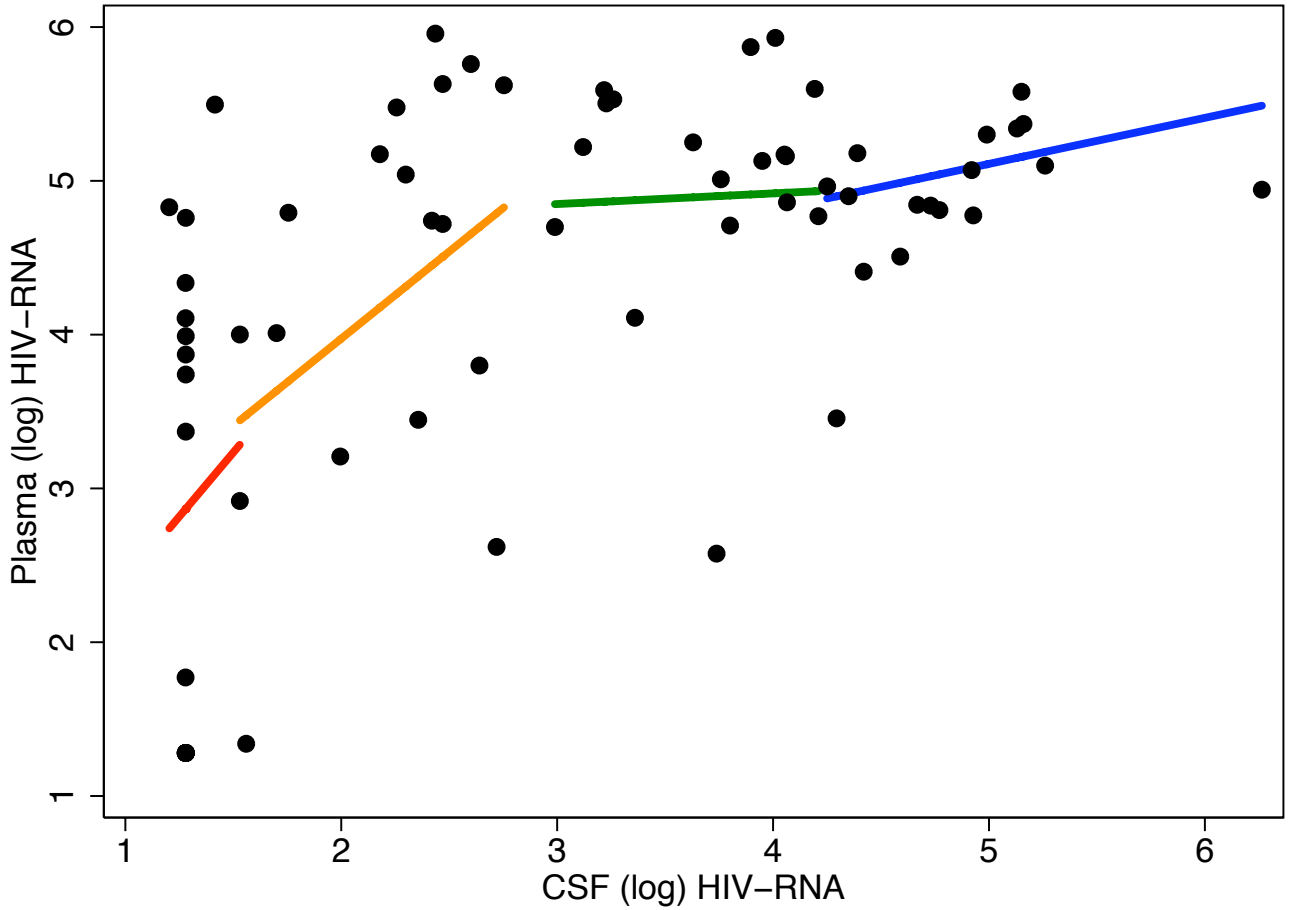
Scatterplot smoother

- Non-parametric method
- Connects a series of *locally fitted* lines
- Result: a flexible smooth curve
- Estimates average value of outcome as a function of predictor, without assuming that the relationship is linear
- Useful for exploring linearity of association

Lowess smoothing in Stata

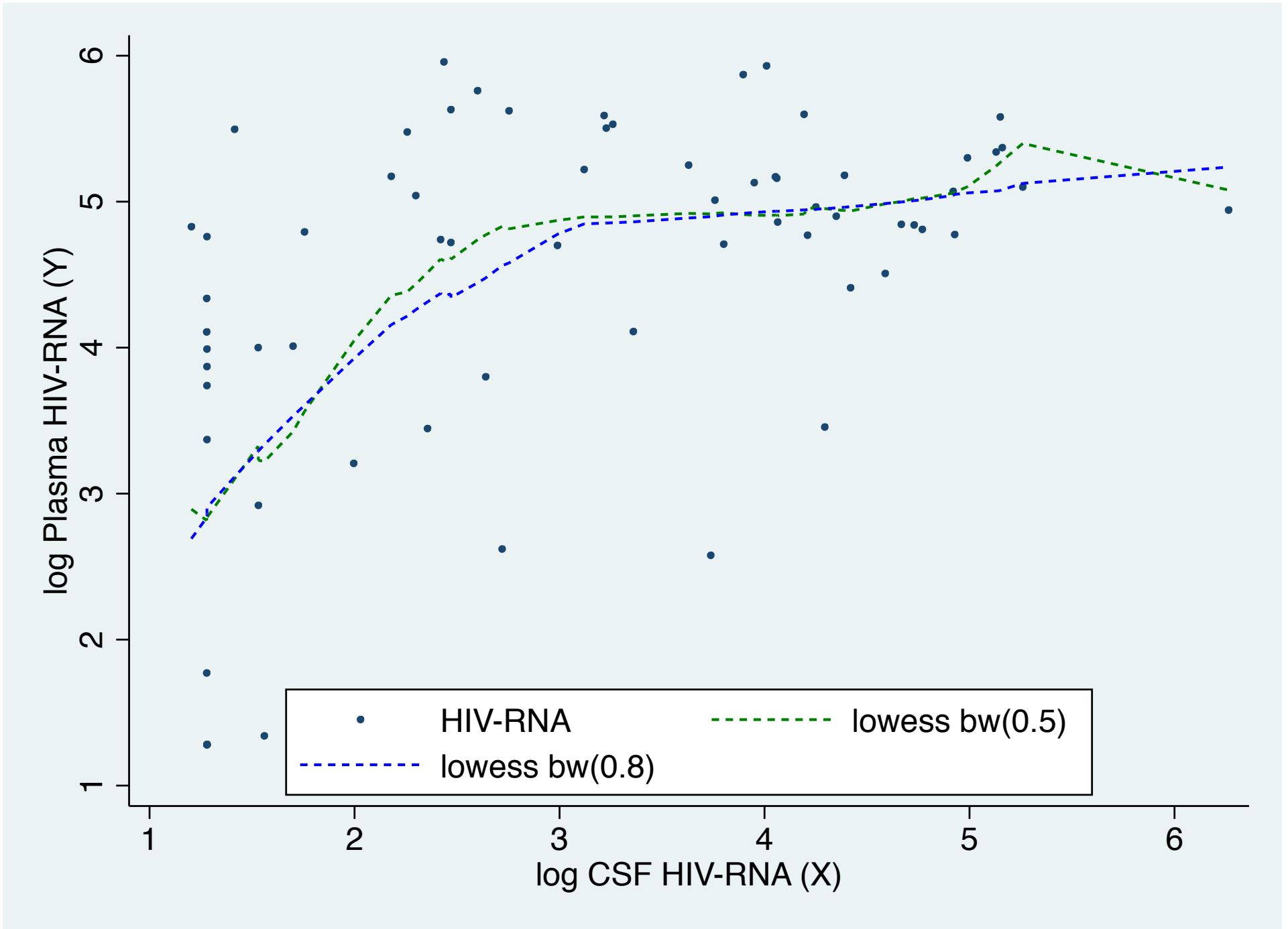
- lowess : *lo*cally *we*ighted *s*catterplot *s*moother
- `lowess hivrna csfrna, bw(0.5)`
- Requests that Stata produce a scatterplot of `hivrna` vs. `csfrna` , including a lowess smooth of the relationship between average outcome and predictor levels
- The `bw( )` (bandwidth) option controls degree of smoothness

Lowess smoothing



- `lowess` joins estimates from lines fitted in local "neighborhoods" of the predictor into a single smooth curve
  - neighborhoods for all predictor values are used (example in figure shows 4)
- The *bandwidth* (`bw`) parameter controls how much data is present in each "neighborhood"

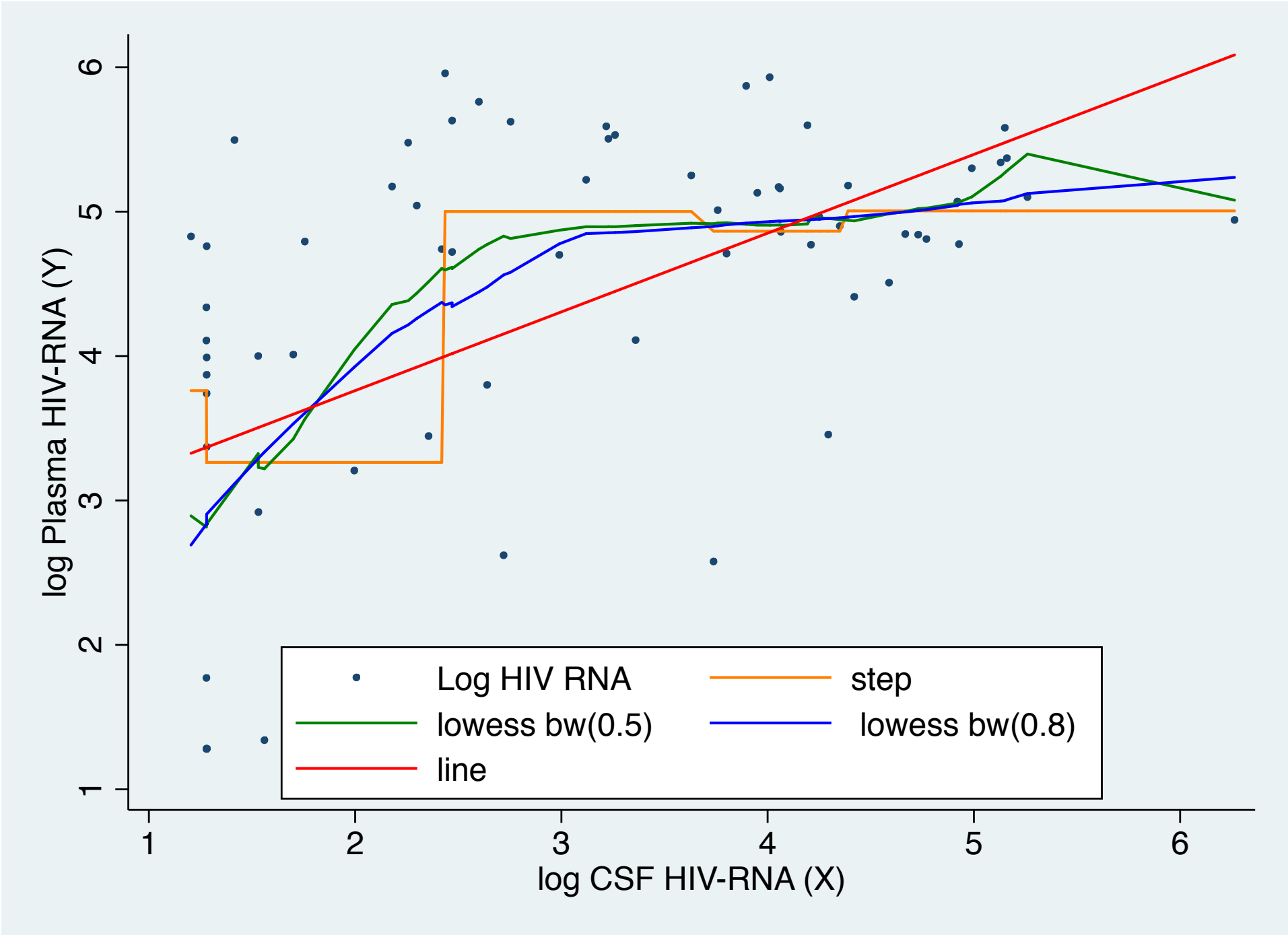
Lowess smooth for two bandwidths



How smooth?

- Smoothness in Lowess increases with bandwidth
  - Too smooth: *imposes more or less linear fit, gives little information about non-linearity and more bias-prone*
  - Not smooth enough: *estimate is sensitive to small variations in the data, obscuring overall trends (i.e. more variable)*
- Larger bandwidth needed in small samples
- Smooths may be unstable at plot margins where data are often lacking

How smooth?


log Plasma HIV-RNA (Y)
1
2
3
4
5
6
log CSF HIV-RNA (X)
1
2
3
4
5
6
Log HIV RNA
step
lowess bw(0.5)
lowess bw(0.8)
line

Definition of multi-predictor linear regression model

- Linear model for a continuous outcome y and p predictors:

  - mean of y is linearly related to the x's:

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

  - individual values of y are linearly related to the x's:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

- β_0 is the mean of y when all x's are 0
- β_j is the change in the mean of y associated with a unit increase in x_j, holding the values of all the other x's fixed
- Coefficients estimated via least squares
- Coefficients for each predictor account for (or "control for") the presence of the other predictors - often called "*adjusted coefficients*"
- Errors (ε) are independent, normal, with 0 mean and constant variance

Linear regression model with two predictors: *Interpretation of coefficients*

- Conditional mean of outcome is linearly related to the predictors:

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

From HERS study data, as discussed in section 4.1 in textbook:

- Example 1: glucose level *(mg/dl)* related to exercise (exercise≥3 times / week,Y=1/N=0), controlling for BMI (kg/m^2) among women without diabetes at baseline

$$\mathrm{E}\left[\mathrm{glucose}|\mathrm{exercise}, \mathrm{BMI}\right] = \beta_0 + \beta_1 \mathrm{exercise} + \beta_2 \mathrm{BMI}$$

- Example 2: glucose level related to HDL cholesterol and BMI (among women without diabetes)

$$\mathrm{E}\left[\mathrm{glucose}|\mathrm{HDL}, \mathrm{BMI}\right] = \beta_0 + \beta_1 \mathrm{HDL} + \beta_2 \mathrm{BMI}$$

- Example 1: Glucose level related to binary indicator of exercise≥ 3 times/wk and BMI

Unadjusted effect of exercise on glucose in HERS (ignoring BMI, and *restricted to non-diabetic women*):

```
. regress glucose i.exercise if diabetes==0
```

```
      Source |       SS       df       MS              Number of obs =    2032
-------------+------------------------------           F(  1,  2030) =   14.97
       Model |  1412.50418     1  1412.50418            Prob > F      =  0.0001
    Residual |  191605.195  2030  94.3867954            R-squared     =  0.0073
-------------+------------------------------           Adj R-squared =  0.0068
       Total |  193017.699  2031  95.0357946            Root MSE      =  9.7153

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    exercise |
         yes |  -1.692789   .4375862    -3.87   0.000    -2.550954   -.8346242
       _cons |   97.36104   .2815138   345.85   0.000     96.80895    97.91313
------------------------------------------------------------------------------
```

Estimates mean glucose for each level of exercise (i.e. the *unadjusted* effect of exercise)

- For non-exercisers (`exercise = 0`)

$$\mathrm{E}\left[\mathrm{glucose}|\mathrm{exercise}=0\right]=\beta_0=97.36104$$

- For exercisers (`exercise = 1`)

$$\mathrm{E}\left[\mathrm{glucose}|\mathrm{exercise}=1\right]=\beta_0+\beta_1=97.36104\text{ -}1.692789$$

- β_1 Estimates the difference between mean levels in the two groups

Adjusted effect of exercise on glucose (controlling for BMI):

```
. regress glucose i.exercise BMI if diabetes==0

      Source |       SS       df       MS              Number of obs =    2030
-------------+------------------------------           F(  2,  2027) =   74.14
       Model |   13153.784     2  6576.89199            Prob > F      =  0.0000
    Residual |  179802.433  2027  88.7037162            R-squared     =  0.0682
-------------+------------------------------           Adj R-squared =  0.0673
       Total |  192956.217  2029  95.0991704            Root MSE      =  9.4183

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    exercise |
         yes |  -.9172885   .4298059    -2.13   0.033    -1.760196   -.0743811
         BMI |   .4736147   .0411907    11.50   0.000     .3928342    .5543952
       _cons |    83.9422   1.199349    69.99   0.000     81.59012    86.29429
------------------------------------------------------------------------------
```

- Estimated coefficient for exercise reflects adjustment for BMI (compare to unadjusted value of -1.69)
- Depending on our underlying causal assumptions, the observed change in coefficient (towards the null value of zero) may reflect the confounding/mediating influence of BMI on average glucose.

Adjusted effect of exercise on glucose (controlling for BMI):

```
. regress glucose i.exercise BMI if diabetes==0

      Source |       SS       df       MS              Number of obs =    2030
-------------+------------------------------           F(  2,  2027) =   74.14
       Model |   13153.784     2  6576.89199            Prob > F      =  0.0000
    Residual |  179802.433  2027  88.7037162            R-squared     =  0.0682
-------------+------------------------------           Adj R-squared =  0.0673
       Total |  192956.217  2029  95.0991704            Root MSE      =  9.4183

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    exercise |
         yes |  -.9172885   .4298059    -2.13   0.033    -1.760196   -.0743811
         BMI |   .4736147   .0411907    11.50   0.000     .3928342    .5543952
       _cons |    83.9422   1.199349    69.99   0.000     81.59012    86.29429
------------------------------------------------------------------------------
```

- *t*-statistic and 95% CI for exercise address the null hypothesis that there is no difference in average glucose between exercise groups, in a model controlling for the effect of BMI as a linear predictor

- Default *F*-statistic evaluates the null hypothesis that the coefficients for exercise & BMI are zero (if not rejected, implies that these variables don't improve on the overall mean glucose level in predicting individual glucose values).

For any fixed value of BMI, β_1 gives the estimated difference in mean glucose between exercisers and non-exercisers:

$$\mathrm{E}\left[\text{glucose}|\text{exercise}{=}1, \text{BMI}\right] = \beta_0 + \beta_1 + \beta_2 \text{BMI} = 83.9422 \text{ -}0.9172885 + 0.4736147 \text{ BMI}$$

$$\mathrm{E}\left[\text{glucose}|\text{exercise}{=}0, \text{BMI}\right] = \beta_0 + \beta_2 \text{BMI} = 83.9422 + 0.4736147 \text{ BMI}$$

For each fixed value of exercise, β_2 gives the estimated increase in mean glucose for a unit increase in BMI:

$$\mathrm{E}\left[\text{glucose}|\text{exercise}{=}1, \text{BMI}\right] = \beta_0 + \beta_1 + \beta_2 \text{BMI} = 83.9422 \text{ -}0.9172885 + 0.4736147 \text{ BMI}$$

$$\mathrm{E}\left[\text{glucose}|\text{exercise}{=}0, \text{BMI}\right] = \beta_0 + \beta_2 \text{BMI} = 83.9422 + 0.4736147 \text{ BMI}$$

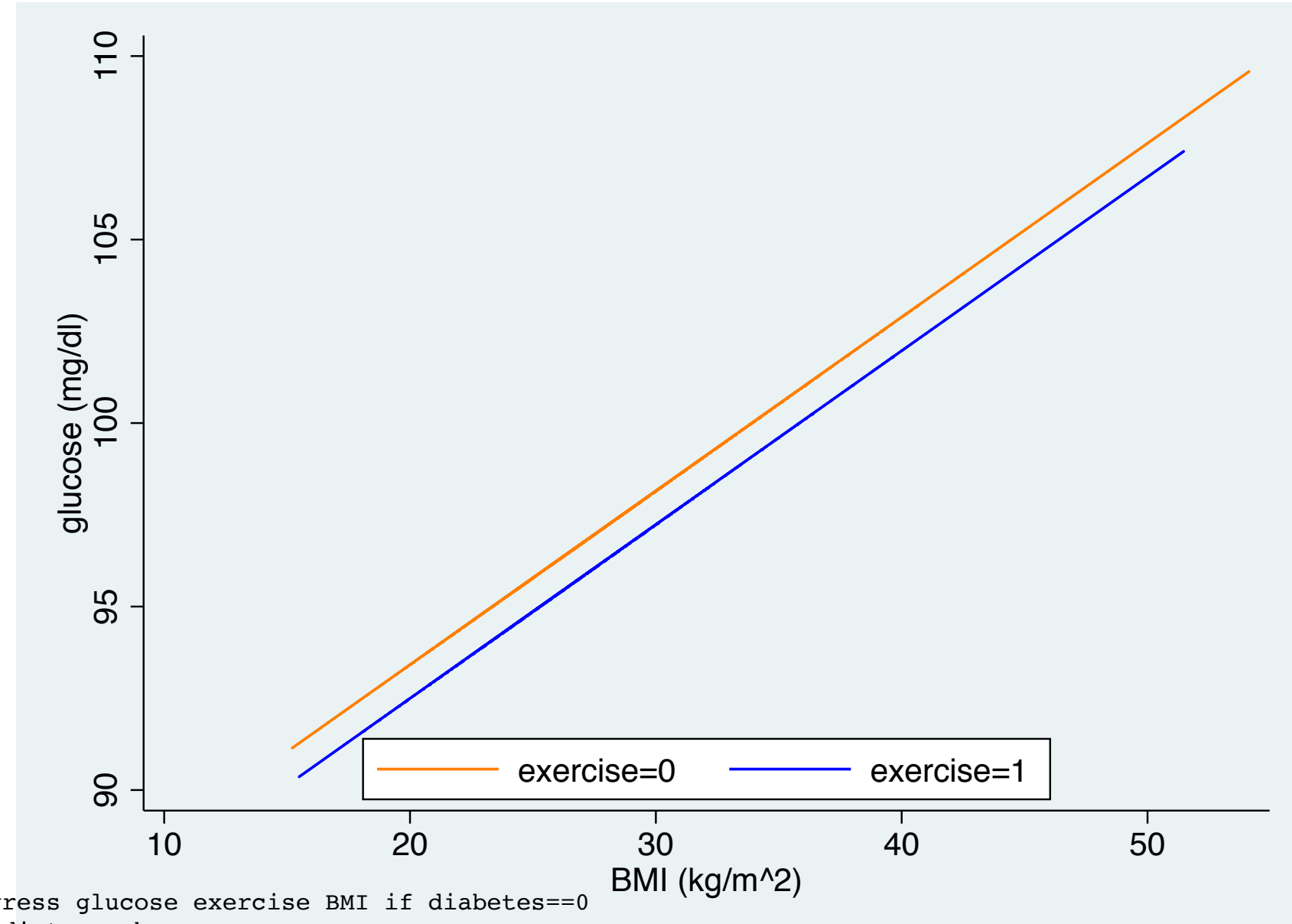
- Specifies a linear relationship between glucose and BMI for each level of exercise

Note: the two lines have identical slopes, but different intercept coefficients . A model allowing different intercepts *and* slopes would represent an *interaction* between BMI and exercise

Each coefficient is *adjusted* for the presence of the other predictor in the model.

Note: For models including complicated functions (e.g. products) of multiple predictors, coefficient interpretations are more complex

- Example 1: average glucose level related to BMI for two levels of exercise




```
. regress glucose exercise BMI if diabetes==0
. predict p, xb

. twoway (line p BMI if exercise==0, sort clpat(solid) lcolor(orange)) (line p BMI if exercise==1, sort
clpat(solid) lcolor(blue)), plotregion(style(none)) ytitle("glucose (mg/dl)") xtitle("BMI (kg/m^2)")
legend(order(1 2) label(1 "exercise=0") label(2 "exercise=1") pos(6) ring(0)) saving(gplot, replace)
graph export gplot.pdf, replace
```

Note on variability of adjusted regression coefficients (sec. 4.2.2.2 in book)

The estimated variability of an adjusted coefficient (e.g. the j_{th}) depends on other variables in the model.

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2_{y|\mathbf{x}}}{(n-1)\sigma^2_{x_j}(1-r^2_j)}$$

$\sigma^2_{y|\mathbf{x}}$ overall residual variance (i.e. variability in y not explained by predictors)

$\sigma^2_{x_j}$ variance of the j_{th} predictor

r^2_j partial correlation between j_{th} and other predictors

- Variability increases with unexplained (residual) variance and with increasing correlation with other predictors
- Variability decreases with increasing sample size, with increased variability in the predictor, and with decreasing correlation with other predictors
- $\frac{1}{(1-{r_j}^2)}$ is known as the *variance inflation factor* for the j_{th} predictor

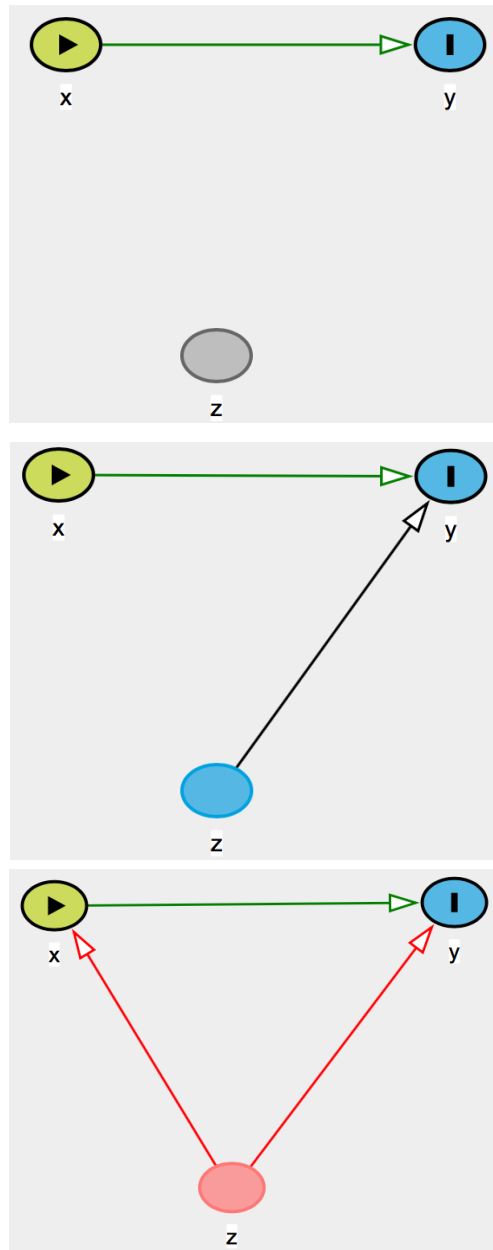
Simulation illustration of the effect of regression adjustment

Consider the relationship between a continuous outcome y with a binary predictor x adjusted for a second continuous predictor z under 3 scenarios:

(1) z unassociated with both x and y

(2) z associated with y, but *not* associated with x

(3) z associated with both x and y



Variables are generated as follows:

z ~ normal(mean=10, SD=2)

x ~ binary (0/1), with outcome probability P (mean) dependent on z:

$$\log\left[\frac{P(z)}{1-P(z)}\right] = \gamma_0 + \gamma_1 z$$

y related to x and z via the linear regression (ε~normal(mean=0, SD=2)):

$$y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon$$

Note: dependence between x and z controlled by γ_1 ; dependence between y and z controlled by β_2 ; True value of $\beta_1 = 1$

```
* Simulation code (takes γ1 and β2 as arguments)
program define simreg
args n gamma1 beta2
drop _all
set obs `n'
* generate a standard normal random variable z (mean 10, SD 2)
gen z = rnormal(10,2)
* generate a binomial random variable x with specified dependence on z
gen px = 1/(1 + exp(-(`gamma1'*z)))
gen x = rbinomial(1,px)
* generate normal error & y - true adjusted coef for x is 1
gen eps = rnormal(0,2)
gen y = 1 + 1*x + `beta2'*z + eps
* fit unadjusted model
reg y x
matrix sb = r(table)
* extract coefficient & SD
scalar mbeta1 = sb[1,1]
scalar sdmbeta1 = sb[2,1]
* fit adjusted model
reg y x z
* extract coefficients & SDs
matrix rb = r(table)
scalar beta1 = rb[1,1]
scalar beta2 = rb[1,2]
scalar sdbeta1 = rb[2,1]
scalar sdbeta2 = rb[2,2]
end
```

In Stata (1000 runs for each scenario outlined above):

```
set seed 1162018
* scenario 1 (z not associated with either x or y)
simulate mbeta1=mbeta1 sdmbeta1=sdmbeta1 beta1=beta1 sdbeta1=sdbeta1, reps(1000): simreg 1000 0 0
* scenario 2 (z associated with y, but not x)
simulate mbeta1=mbeta1 sdmbeta1=sdmbeta1 beta1=beta1 sdbeta1=sdbeta1, reps(1000): simreg 1000 0 1
* scenario 3 (z associated with both x or y)
simulate mbeta1=mbeta1 sdmbeta1=sdmbeta1 beta1=beta1 sdbeta1=sdbeta1, reps(1000): simreg 1000 0.25 1
```

Simulation illustration of the effect of regression adjustment

Results for effect of x (β_1) in all 3 scenarios (1000 simulations each):

(1) z unassociated with both x and y: $\gamma_1 = 0$; $\beta_2 = 0$

(2) z associated with y, but *not* associated with x: $\gamma_1 = 0$; $\beta_2 = 1$

(3) z associated with both x and y: $\gamma_1 = 0.25$; $\beta_2 = 1$

Scenario	*Unadjusted* β_1 (SE)	*Adjusted* β_1 (SE)
(1)	1.000 (0.127)	1.000 (0.127)
(2)	1.009 (0.179)	1.005 (0.127)
(3)	1.994 (0.329)	1.006 (0.232)

Conclusions for each scenario:

(1) No obvious effect of adjustment on coefficient or variability

(2) Adjustment does not affect estimate, but decreases variability (implications for RCTs)

(3) Adjustment affects estimate and variability (reflects confounding effect of z)

Example: Binary Predictors with Reversed Categories

- Define a new exercise variable as follows:
  - 1 : exercise < 3 times / week ; 0 : exercise ≥ 3 times / week
- Regression model (next slide) still specifies mean glucose for both levels of exercise
- What will intercept and slope be for this model?

Example: Binary Predictor with Reversed Categories

```
. recode exercise (1 = 0) (0 = 1), gen(nexercise)
(2763 differences between exercise and nexercise)

. regress glucose i.nexercise BMI if diabetes==0

      Source |       SS       df       MS              Number of obs =    2030
-------------+------------------------------           F(  2,  2027) =   74.14
       Model |   13153.784     2  6576.89199            Prob > F      =  0.0000
    Residual |  179802.433  2027  88.7037162            R-squared     =  0.0682
-------------+------------------------------           Adj R-squared =  0.0673
       Total |  192956.217  2029  95.0991704            Root MSE      =  9.4183

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 1.nexercise |   .9172885   .4298059     2.13   0.033     .0743811    1.760196
         BMI |   .4736147   .0411907    11.50   0.000     .3928342    .5543952
       _cons |   83.02491    1.14656    72.41   0.000     80.77635    85.27347
------------------------------------------------------------------------------
```

Example: Binary Predictor with Reversed Categories

Stata syntax for changing reference category for a categorical predictor:

```
. regress glucose ib1.exercise BMI if diabetes==0

      Source |       SS       df       MS              Number of obs =    2030
-------------+------------------------------           F(  2,  2027) =   74.14
       Model |   13153.784     2  6576.89199            Prob > F      =  0.0000
    Residual |  179802.433  2027  88.7037162            R-squared     =  0.0682
-------------+------------------------------           Adj R-squared =  0.0673
       Total |  192956.217  2029  95.0991704            Root MSE      =  9.4183

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    exercise |
          no |   .9172885   .4298059     2.13   0.033     .0743811    1.760196
         BMI |   .4736147   .0411907    11.50   0.000     .3928342    .5543952
       _cons |   83.02491    1.14656    72.41   0.000     80.77635    85.27347
------------------------------------------------------------------------------
```

Note: the "b." syntax for specifying base levels of categorical predictors is covered in section 11.4.3.2 of the Stata manual:

- Example 2: average glucose level related to 2 continuous predictors (BMI and HDL)

```
. regress glucose BMI HDL if diabetes==0

      Source |       SS       df       MS              Number of obs =    2023
-------------+------------------------------           F(  2,  2020) =   74.50
       Model |  13237.1205     2  6618.56023            Prob > F      =  0.0000
    Residual |  179454.635  2020  88.8389284            R-squared     =  0.0687
-------------+------------------------------           Adj R-squared =  0.0678
       Total |  192691.756  2022  95.2976043           Root MSE      =  9.4254

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         BMI |    .468405   .0413542    11.33   0.000     .3873038    .5495062
         HDL |  -.0393889   .0157441    -2.50   0.012    -.0702653   -.0085125
       _cons |   85.73492   1.527376    56.13   0.000     82.73952    88.73031
------------------------------------------------------------------------------
```

Specifies a linear relationship between glucose and BMI for any fixed value of HDL

- example: HDL= 100:

$$\mathrm{E}\left[\mathrm{glucose}|\mathrm{BMI}, \mathrm{HDL}{=}100\right] = \beta_0 + \beta_1 \mathrm{BMI} + \beta_2 100 = 85.73492 - 3.93889 + 0.468405\ \mathrm{BMI}$$

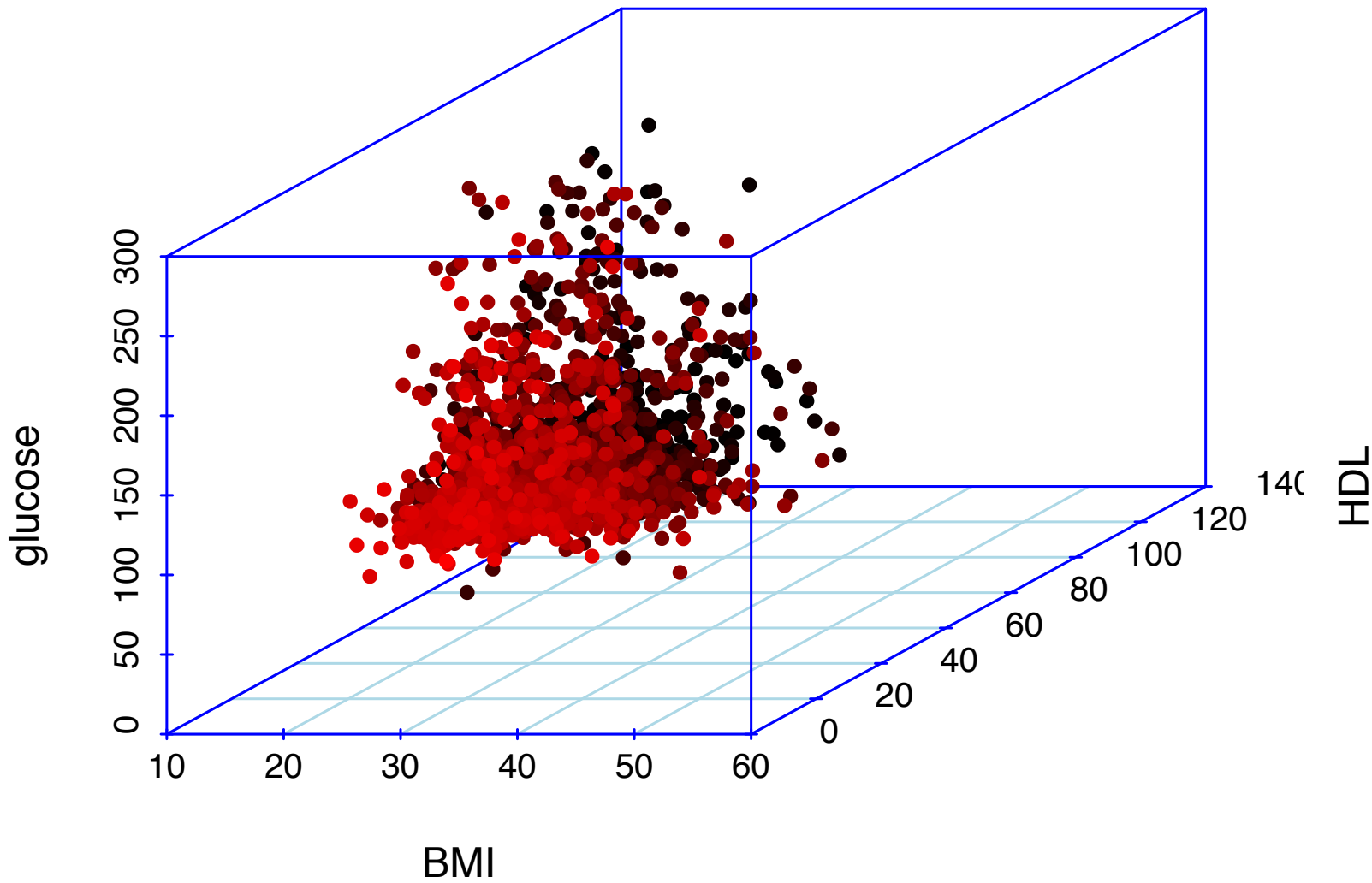
Specifies a linear relationship between glucose and HDL for any fixed value of BMI

- example: BMI=50:

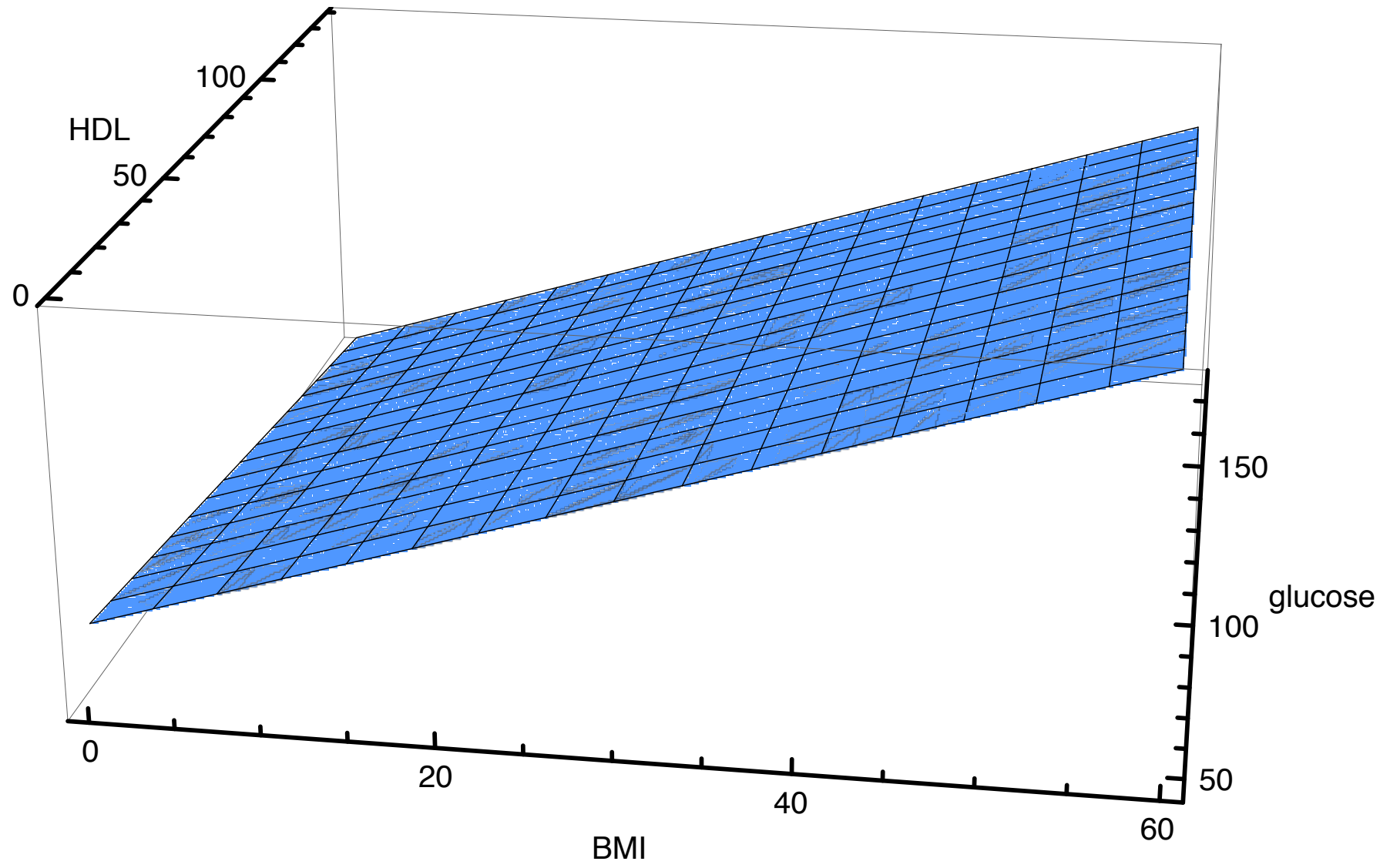
$$\mathrm{E}\left[\mathrm{glucose}|\mathrm{BMI}{=}50, \mathrm{HDL}\right] = \beta_0 + \beta_1 50 + \beta_2 \mathrm{HDL} = 85.73492 + 0.468405 \times 50 - 0.0393889\ \mathrm{HDL}$$

Note: different fixed values of one predictor generate parallel lines for the relationship between the outcome and the other predictor

- Example 2: scatter plot for glucose level related to BMI and HDL



- Example 2: linear model for average glucose level related to BMI and HDL



Centering in Multiple Predictor Models

- Example : center BMI and HDL

```
. egen meanBMI = mean(BMI) if diabetes==0

. gen cBMI = BMI - meanBMI
(5 missing values generated)

. egen meanHDL = mean(HDL) if diabetes==0

. gen cHDL = HDL - meanHDL
(11 missing values generated)

. regress glucose cBMI cHDL

      Source |       SS       df       MS              Number of obs =    2023
-------------+------------------------------           F(  2,  2020) =   74.50
       Model |  13237.1205     2  6618.56023            Prob > F      =  0.0000
    Residual |  179454.635  2020   88.8389284           R-squared     =  0.0687
-------------+------------------------------           Adj R-squared =  0.0678
       Total |  192691.756  2022   95.2976043           Root MSE      =  9.4254

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cBMI |    .468405   .0413542    11.33   0.000     .3873038    .5495062
        cHDL |  -.0393889   .0157441    -2.50   0.012    -.0702653   -.0085125
       _cons |   96.66618   .2095579   461.29   0.000     96.25521    97.07715
------------------------------------------------------------------------------
```

Coefficients for BMI and HDL unchanged from previous model

Intercept gives mean glucose for individuals with mean values of BMI and HDL

Centering Predictors : Summary

- 1-unit change in BMI = 1-unit change in centered BMI
- 1-unit change in HDL = 1-unit change in centered HDL
- Intercept is now mean glucose for an individual with average HDL & average BMI
- Centering makes intercept interpretable
- Centering also helps reduce collinearity in models with interaction terms

Rescaling Predictors

- Example : rescale BMI by 10 and HDL by 50

```
. gen BMI_10 = BMI/10
(5 missing values generated)

. gen HDL_50 = HDL/50
(11 missing values generated)

. regress glucose BMI_10 HDL_50 if diabetes==0

      Source |       SS       df       MS              Number of obs =    2023
-------------+------------------------------           F(  2,  2020) =   74.50
       Model |  13237.1203     2  6618.56017            Prob > F      =  0.0000
    Residual |  179454.635  2020  88.8389284            R-squared     =  0.0687
-------------+------------------------------           Adj R-squared =  0.0678
       Total |  192691.756  2022  95.2976043           Root MSE      =  9.4254

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      BMI_10 |    4.68405   .4135415    11.33   0.000     3.873038    5.495062
      HDL_50 |  -1.969444   .7872055    -2.50   0.012    -3.513264   -.4256248
       _cons |   85.73492   1.527376    56.13   0.000     82.73952    88.73031
------------------------------------------------------------------------------
```

coefficient gives change mean glucose for each 10-unit increase in BMI

coefficient gives change mean glucose for each 50-unit increase in HDL

intercept gives mean glucose for an indiv. with BMI = 0 & HDL = 0

Rescaling Predictors: Summary

- slopes for rescaled versions of BMI & HDL change to represent scaling units
- leads to estimated coefficients with clearer clinical interpretations
- coefficient for `BMI_10` = 10*coef. for BMI , coef. for `HDL_50` = 50*coef. for HDL
- CIs for coefficients are also multiplied by 10 (for BMI) and 50 (for HDL), respectively
- Hypothesis tests for model and coefficients, R^2 unaffected
- The same model -- expressed differently

Centering & Rescaling Predictors

- Combines the effects if centering and rescaling
- Example : rescale centered BMI by 10 and centered HDL by 50

```
. gen cBMI_10 = cBMI/10
(11 missing values generated)

. gen cHDL_50 = cHDL/50
(11 missing values generated)

. regress glucose cBMI_10 cHDL_50, if diabetes==0

      Source |       SS       df       MS              Number of obs =    2023
-------------+------------------------------           F(  2,  2020) =   74.50
       Model |  13237.1204     2  6618.56022            Prob > F      =  0.0000
    Residual |  179454.635  2020  88.8389284            R-squared     =  0.0687
-------------+------------------------------           Adj R-squared =  0.0678
       Total |  192691.756  2022  95.2976043           Root MSE      =  9.4254

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     cBMI_10 |     4.68405   .4135415    11.33   0.000     3.873038    5.495062
     cHDL_50 |   -1.969444   .7872055    -2.50   0.012    -3.513264   -.4256248
       _cons |    96.66618   .2095579   461.29   0.000     96.25521    97.07715
------------------------------------------------------------------------------
```

coefficient gives change mean glucose for each 10-unit increase in BMI

coefficient gives change mean glucose for each 50-unit increase in HDL

intercept gives mean glucose for an indiv. with average BMI & average HDL

Standardized Regression Coefficients

- Centering, and scaling by the SD of predictor *and* outcome leads to "standardized" regression coefficients
- Standardized coefficients for predictors have a uniform interpretation:
  - measure change in mean outcome (in SD units) for a one SD increase in predictor
- Can aid in assessing strength of assoc. for predictors in multiple predictor models
- Most useful for continuous predictors

standardized regression coefficients with Stata

```
. regress glucose BMI cHDL if diabetes==0, beta

      Source |       SS       df       MS              Number of obs =    2023
-------------+------------------------------           F(  2,  2020) =   74.50
       Model |  13237.1205     2  6618.56023            Prob > F      =  0.0000
    Residual |  179454.635  2020  88.8389284            R-squared     =  0.0687
-------------+------------------------------           Adj R-squared =  0.0678
       Total |  192691.756  2022  95.2976043           Root MSE      =  9.4254

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
         BMI |    .468405   .0413542    11.33   0.000                 .2470035
        cHDL |  -.0393889   .0157441    -2.50   0.012                -.0545577
       _cons |   83.70747   1.162874    71.98   0.000                         .
------------------------------------------------------------------------------
```

- `beta` option request that standardized coeff. are reported in place of 95% confidence intervals
- this option *standardizes both outcome & predictors*

Centering and Rescaling, Summary:

- Center variables: affects interpretation of intercept
- Rescale variables: affects interpretation of slope
- Centering and rescaling: combines the above effects in one model
  - *standardized regression* is a special case (results often less interpretable clinically)

Post estimation using regression models

- Used to estimate quantities of interest based on a previously fitted model
  - *predicted average outcomes for particular predictor values*
  - *marginal mean estimates*
  - *help in interpreting models with interactions*

Example: `lincom` to estimate outcome means for specified predictor values

- Example: conditional means for `exercise = 1` & `exercise =0`, for `BMI=20`
- The `lincom` command is issued immediately after a regression model is fitted
- Gives the estimated mean outcome (and 95% CI) using a linear combination of coefficients

$$\mathrm{E}\left[\text{glucose}|\text{exercise} = 1, \text{BMI} = 20\right] = \beta_0 + \beta_1 + \beta_2 20$$

$$\mathrm{E}\left[\text{glucose}|\text{exercise} = 0, \text{BMI} = 20\right] = \beta_0 + \beta_2 20$$

```
. quietly regress glucose i.exercise BMI if diabetes==0
. lincom _cons + 1.exercise + BMI*20

 ( 1)  1.exercise + 20*BMI + _cons = 0
------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   92.49721   .4260675   217.10   0.000     91.66163    93.33278
------------------------------------------------------------------------------

. lincom _cons + BMI*20

 ( 1)  20*BMI + _cons = 0
------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    93.4145   .4392638   212.66   0.000     92.55304    94.27595
------------------------------------------------------------------------------
```

prefacing command by "quietly" requests that results are stored but not printed

$$\mathrm{E}\left[\text{glucose}|\text{exercise}=1,\text{BMI}=20\right]-\mathrm{E}\left[\text{glucose}|\text{exercise}=0,\text{BMI}=20\right]=\beta_1$$

```
. lincom exercise

 ( 1)  exercise = 0

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.9172885   .4298059    -2.13   0.033    -1.760196   -.0743811
------------------------------------------------------------------------------
```

- The estimated regression coefficient $\beta_1 = -0.9173$ estimates the conditional effect of exercise on glucose (expressed as a difference) *for any fixed BMI*

Example: extracting regression coefficients from a fitted model

- Stata saves regression results until another model is fitted. These can be used for a variety of purposes. Coefficients and std. errors are stored in system variables `_b[]` and `_se[]`

- Example: conditional means for `exercise = 1` & `exercise =0`, for `BMI=20`

```
. quietly regress glucose i.exercise BMI if diabetes==0
. display _b[_cons]
83.942202

. display _b[1.exercise]
-.91728846

. display _b[BMI]
.47361469

. display _b[_cons] + _b[1.exercise] + _b[BMI]*20
92.497207

. display _b[_cons] + _b[BMI]*10
93.414496
```

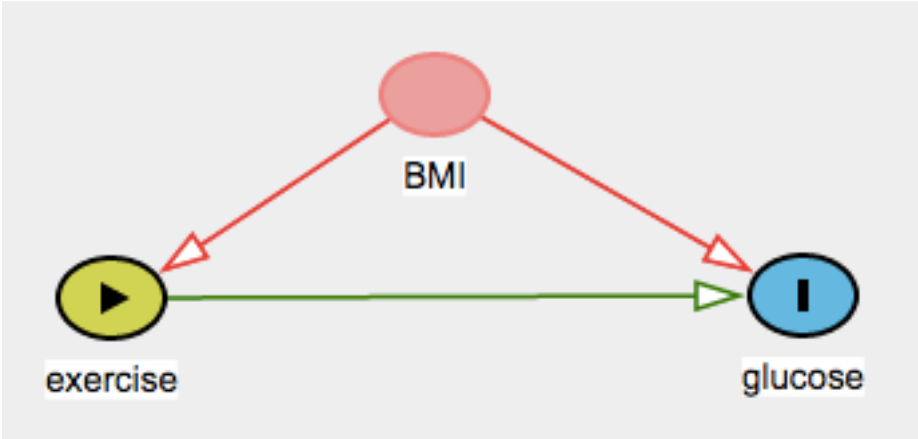
- In this case, `lincom` proves more useful since it also computes 95% CIs

- Note: complete regression results are stored as well, and can be viewed with the command

```
.   ereturn list
```

Regression-based estimate of a *marginal causal effect*

- Assumed (oversimplified) causal model:



- Define *average causal effect* as the difference between average glucose in the population under (possibly counterfactual) "assignment" to exercise and non-exercise:

$$\mathrm{E}\left[\mathrm{glucose}(\mathrm{exercise}{=}1)\right] - \mathrm{E}\left[\mathrm{glucose}(\mathrm{exercise}{=}0)\right]$$

- This difference is a *marginal causal effect* (see Chapt. 9)

- Since each individual is observed at only one exercise level, this must be estimated based on the sample, involving assumptions about the exchangeability (as achieved in randomization) of 'assignment" to exercise.

- As explained above (slide 38), the adjusted effect of exercise given by the regression coefficient is interpreted as "conditional" on the value of BMI. The corresponding marginal effect is not conditional on a particular BMI value, and is interpreted as a population level effect.

Regression-based estimate of a *marginal causal effect*

- Assuming the causal model and assumptions are correct, the marginal causal effect can be estimated by the marginal means in the two exercise groups:

$$\mathrm{E}\left[\mathrm{glucose}|\mathrm{exercise}=1\right]-\mathrm{E}\left[\mathrm{glucose}|\mathrm{exercise}=0\right]$$

- Since the fitted model adjusts for BMI, the marginal means with respect to exercise can be obtained by averaging (a.k.a. "margining") over the distribution of BMI in the sample:

$$\mathrm{E}\left[\mathrm{glucose}|\mathrm{exercise}=x\right]=\mathrm{E}_z\left[\mathrm{E}\left[\mathrm{glucose}|\mathrm{exercise}=x,\mathrm{BMI}=z\right]\right]$$

- Stata allows such marginal effects to be estimated from fitted regression models using the `margins` command (issued after a model is fit)

- `margins` uses the previously fitted model (i.e. in this case adjusting for BMI) to estimate the marginal average glucose for the two values of exercise in the entire sample

Preview: Using linear regression to estimate the marginal causal effect of exercise on glucose

- Using `margins` for post estimation after fitting a regression model:

```
. quietly regress glucose i.exercise BMI if diabetes==0
. margins exercise
```

(*Note:* "`i.`" syntax is required for categorical predictors specified in `margins`)

```
Predictive margins                              Number of obs     =       2030
Model VCE    : OLS

Expression   : Linear prediction, predict()
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    exercise |
          no |   97.04504   .2745954   353.41   0.000     96.50653    97.58356
         yes |   96.12776   .3272173   293.77   0.000     95.48604    96.76947
------------------------------------------------------------------------------
```

- Marginal means for the two values of exercise are obtained from the model by fixing exercise and averaging the resulting estimated outcome values over the entire sample

```
. gen y0 = _b[_cons] + _b[1.exercise]*0 + _b[BMI]*BMI if diabetes==0
(733 missing values generated)

. gen y1 = _b[_cons] + _b[1.exercise]*1 + _b[BMI]*BMI  if diabetes==0
(733 missing values generated)

. summ y0 y1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          y0 |      2030    97.04504    2.434996   91.14588    109.579
          y1 |      2030    96.12776    2.434996   90.22859   108.6617
```

Preview: Estimate the marginal causal effect of exercise on glucose

- The *difference* between the marginal means estimates the average causal effect - this can be obtained using the `dydx` option to `margins`

```
. margins, dydx(exercise)

Average marginal effects                        Number of obs   =       2030
Model VCE    : OLS

Expression   : Linear prediction, predict()
dy/dx w.r.t. : 1.exercise

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    exercise |
         yes |  -.9172885   .4298059    -2.13   0.033    -1.760196   -.0743811
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

- This is *identical to the estimated coefficient* from the fitted regression model

- For linear regression, the marginal causal effect can usually be *estimated directly* by the regression coefficient for x (*not necessarily the case for nonlinear models such as logistic regression*)

- So, regression coefficients from strictly linear models can often be interpreted as *both conditional and marginal* (depending on validity of causal assumptions and model)