

Linear Regression

January 12, 2016
Biostatistics 208

This week:

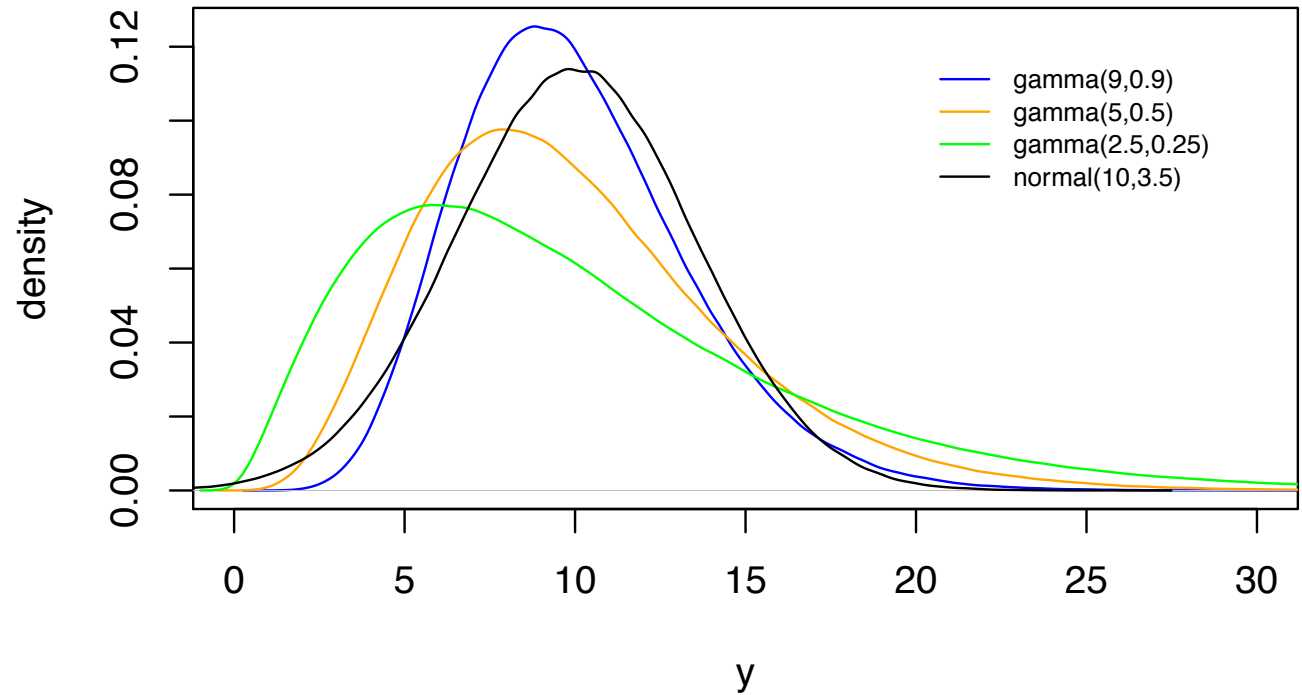
- Review of simple linear regression model
- Centering & scaling predictors for interpretability
- Outliers and assessing linearity in the simple linear model
- Introduction to multiple linear regression
- Post-estimation in multiple predictor models

Review: Tests for non-normality

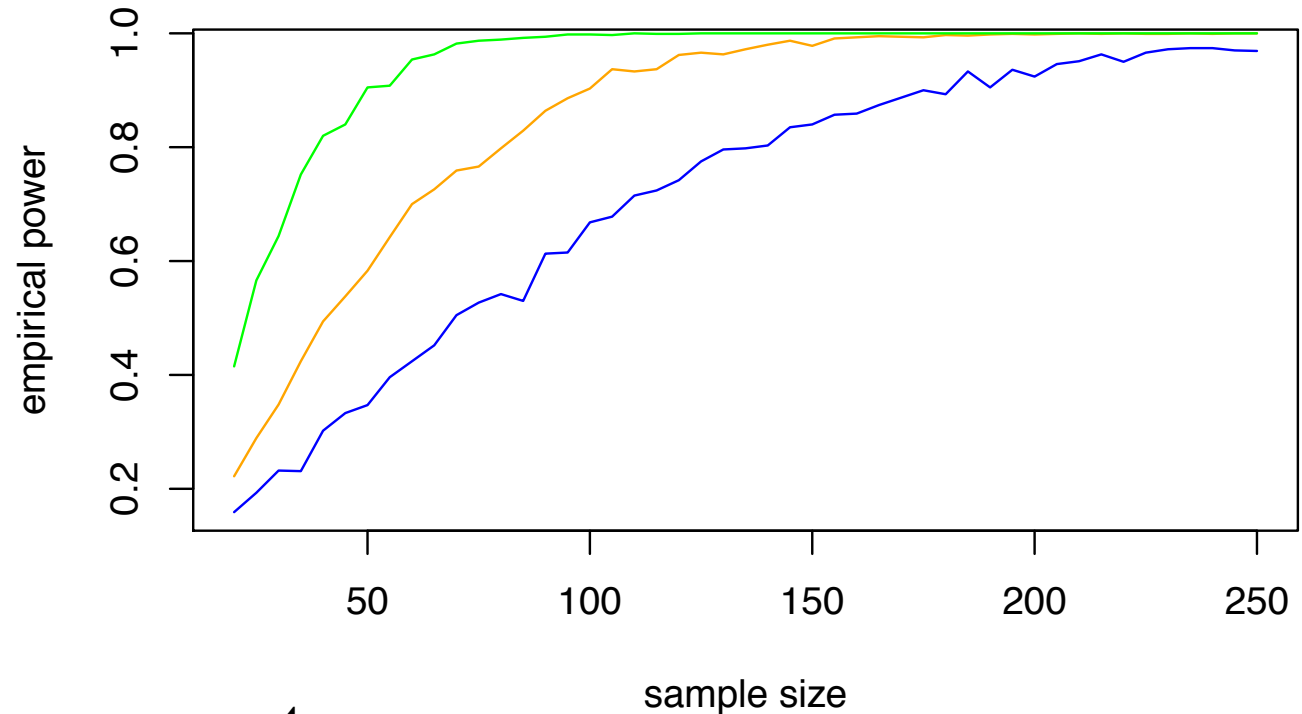
- There are a number of statistical tests for non-normality of a variable (e.g. Shapiro-Wilk, Kolmogorov-Smirnov)
 - null hypothesis: source distribution of the variable is the normal
- Some issues with the testing approach:
 - reject for small deviations from normality in large samples
 - reduced power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

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:
 - 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 test, bootstrap CI), at least as a sensitivity analysis

Simple linear regression model

- Model represents the conditional mean of y as a function of x
 - $E(y | x) = \beta_0 + \beta_1 x$
- 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 variance at all values of x
 - normally distributed (at least in small samples)

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

. regress HDL waist

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

HDL	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
waist	-0.2340937	.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
Model	27749.7008	1	27749.7008
Residual	452355.921	2748	164.612781
Total	480105.622	2749	174.647371

```
Number of obs = 2750
F( 1, 2748) = 168.58
Prob > F      = 0.0000
R-squared     = 0.0578
Adj R-squared = 0.0575
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 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!=. ————— Exclude missing values on outcome  
(11 missing values generated)
```

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

```
. reg HDL cwaist
```

Source	SS	df	MS	Number of obs = 2750		
Model	27749.7008	1	27749.7008	F(1, 2748) = 168.58		
Residual	452355.921	2748	164.612781	Prob > F = 0.0000		
				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

- estimated mean HDL for an indiv. with average waist circumference is 50.27 (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 unless waist size is also centered

Model scaled waist circumference

```
. gen waist_10 = waist/10
```

```
. reg HDL waist_10
```

Source	SS	df	MS	Number of obs = 2750		
Model	27749.7002	1	27749.7002	F(1, 2748) = 168.58		
Residual	452355.922	2748	164.612781	Prob > F = 0.0000		
				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		
Model	27749.7008	1	27749.7008	F(1, 2748) = 168.58		
Residual	452355.921	2748	164.612781	Prob > F = 0.0000		
Total	480105.622	2749	174.647371	R-squared = 0.0578		
				Adj R-squared = 0.0575		
				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		
Model	27749.7008	1	27749.7008	F(1, 2748) = 168.58		
Residual	452355.921	2748	164.612781	Prob > F = 0.0000		
Total	480105.622	2749	174.647371	R-squared = 0.0578		
				Adj R-squared = 0.0575		
				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 mean zero, SD one)
- Note that standardization was restricted to observations with non-missing values of hdl

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

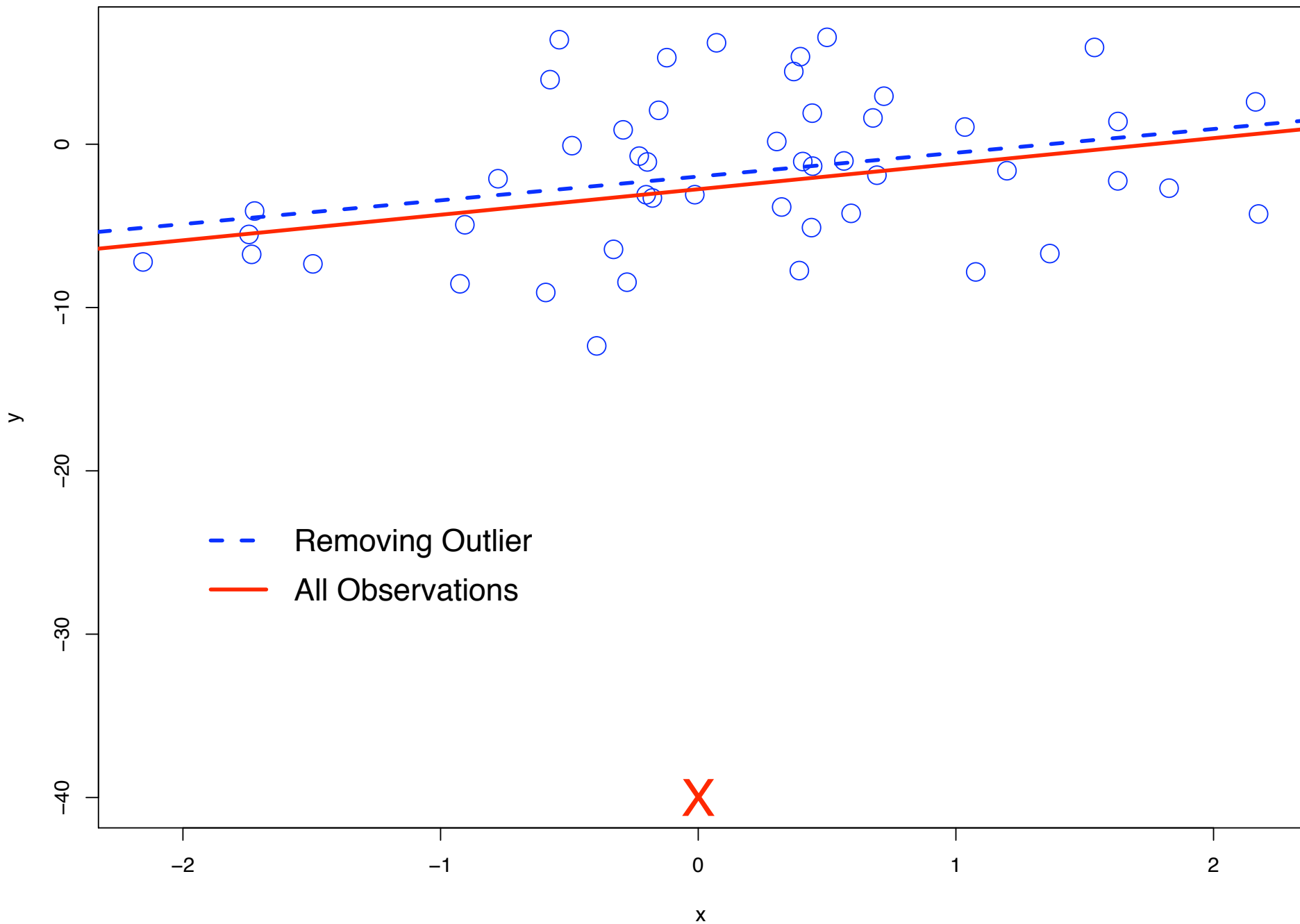
Outliers and influential points

- 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

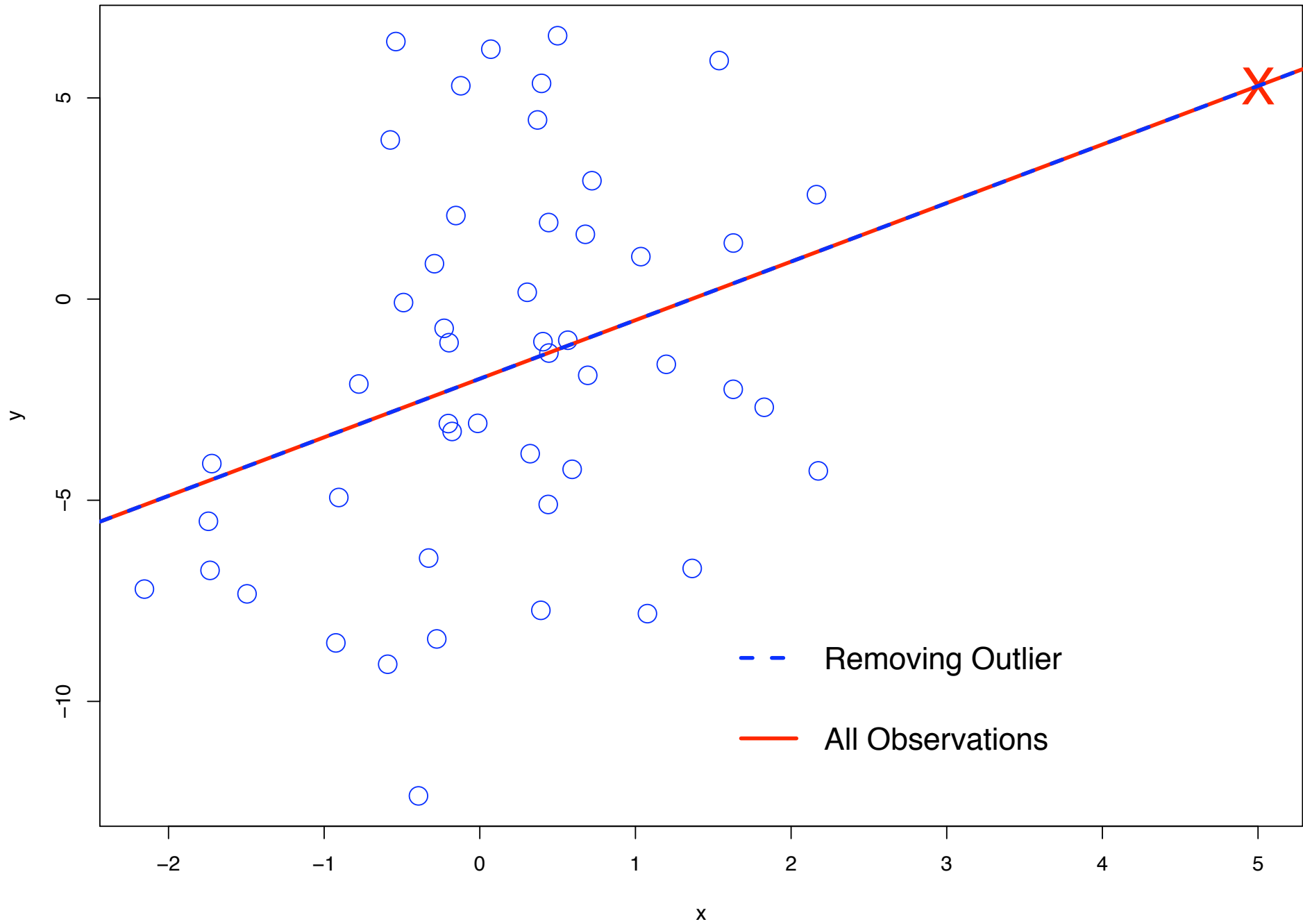
- Benign outlier:
 - *x not an outlier, y far from regression line*
 - *influences the intercept but not the slope*
- High leverage point:
 - *x an outlier, y near regression line*
 - *influences neither slope nor intercept*
 - *may affect precision of estimated coefficients and R^2*
- Influential point:
 - *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

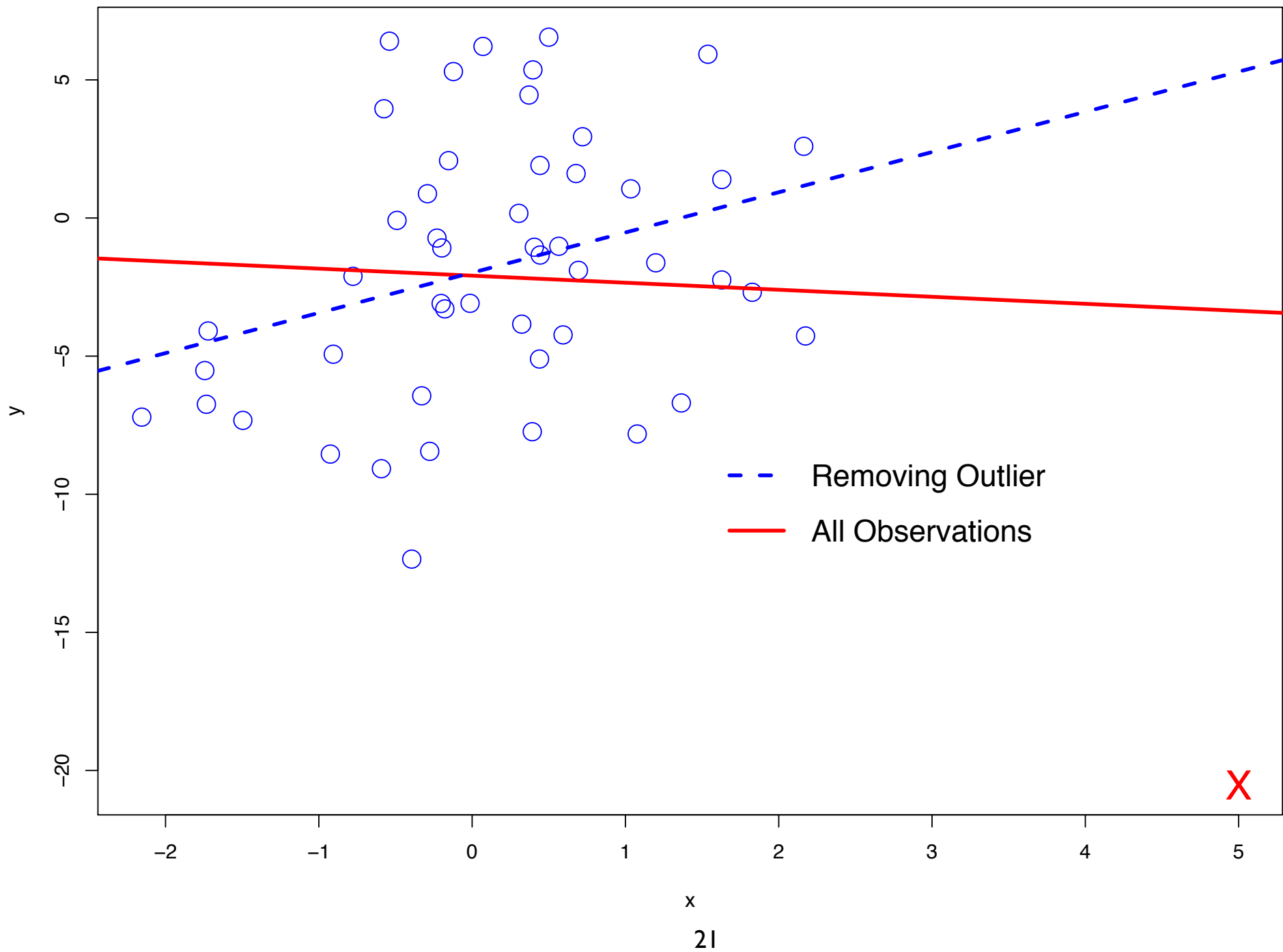


X

Example: High leverage point



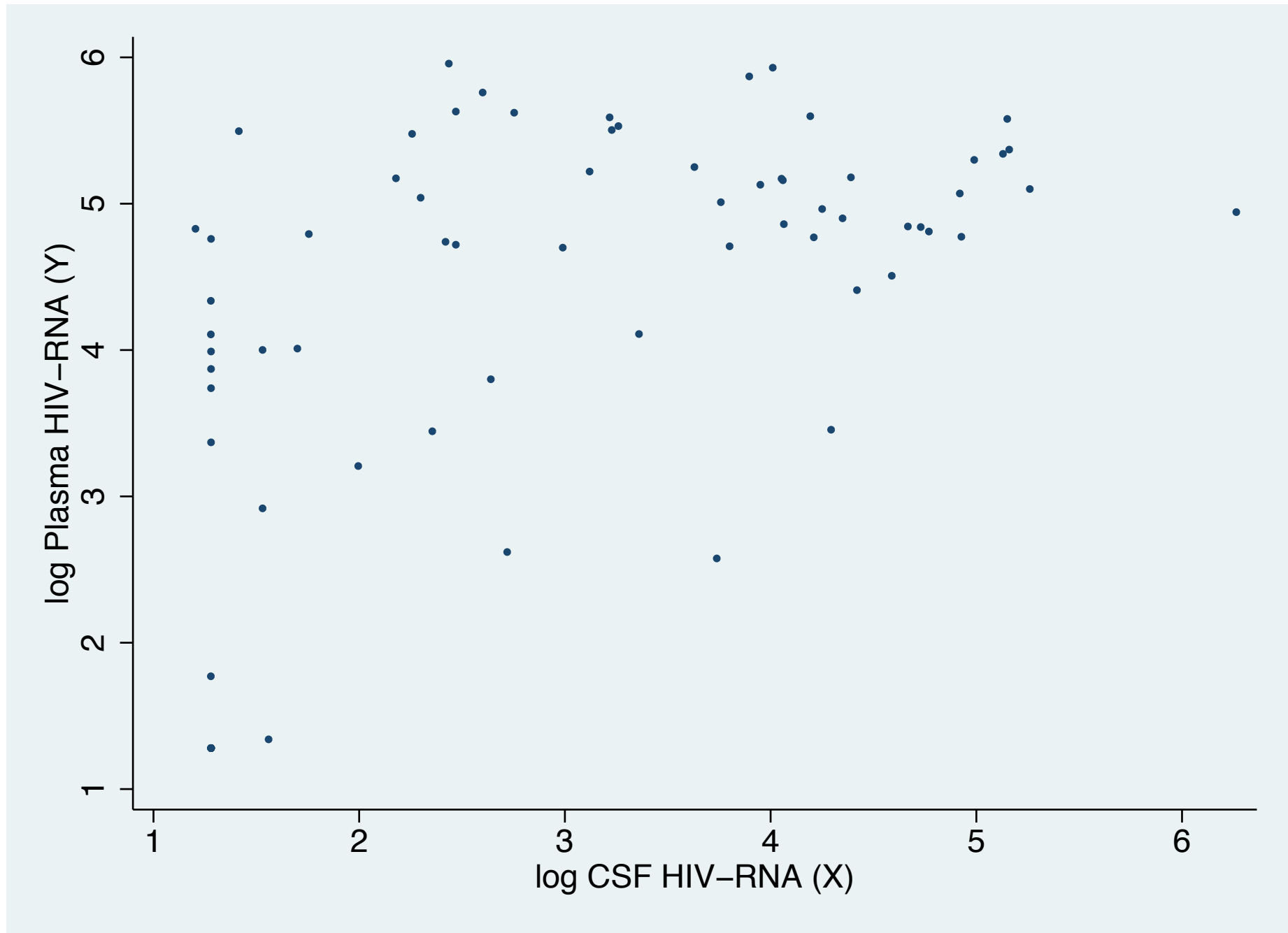
Example: Influential point



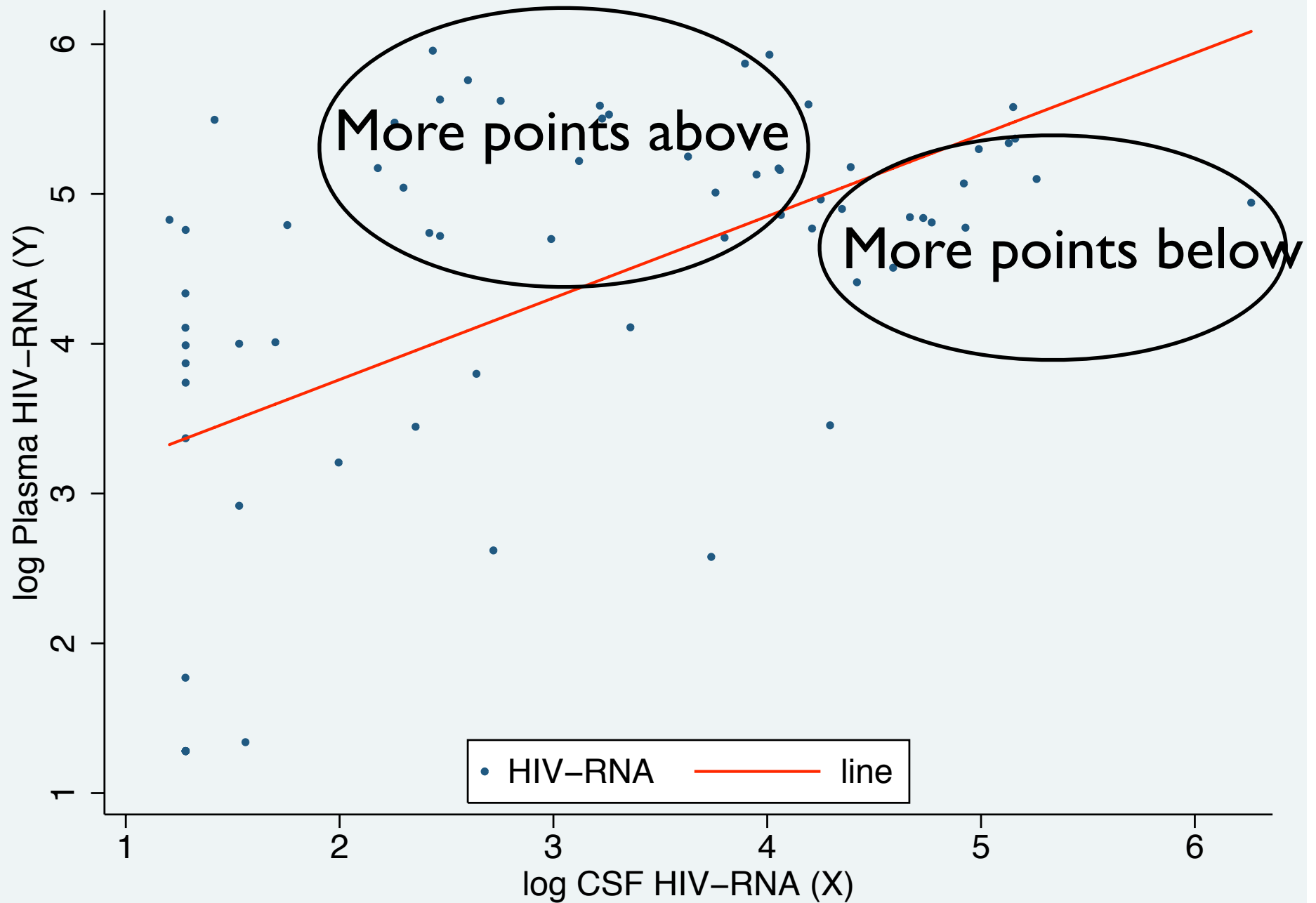
Exploratory methods for assessing linearity

- 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 (future lecture topic)
- 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?



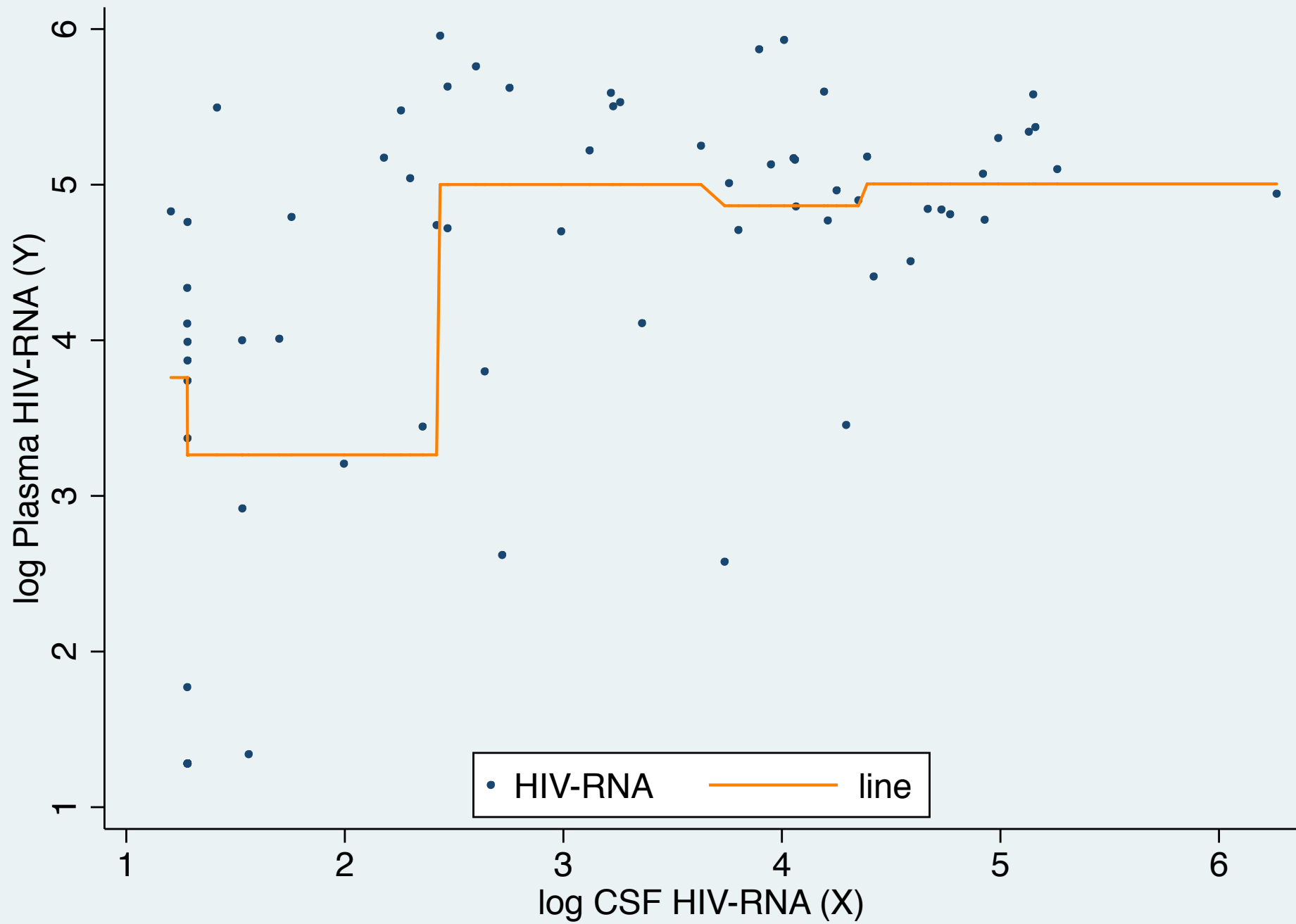
Step Function

- Split continuous predictor into several equally spaced categories
- 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)

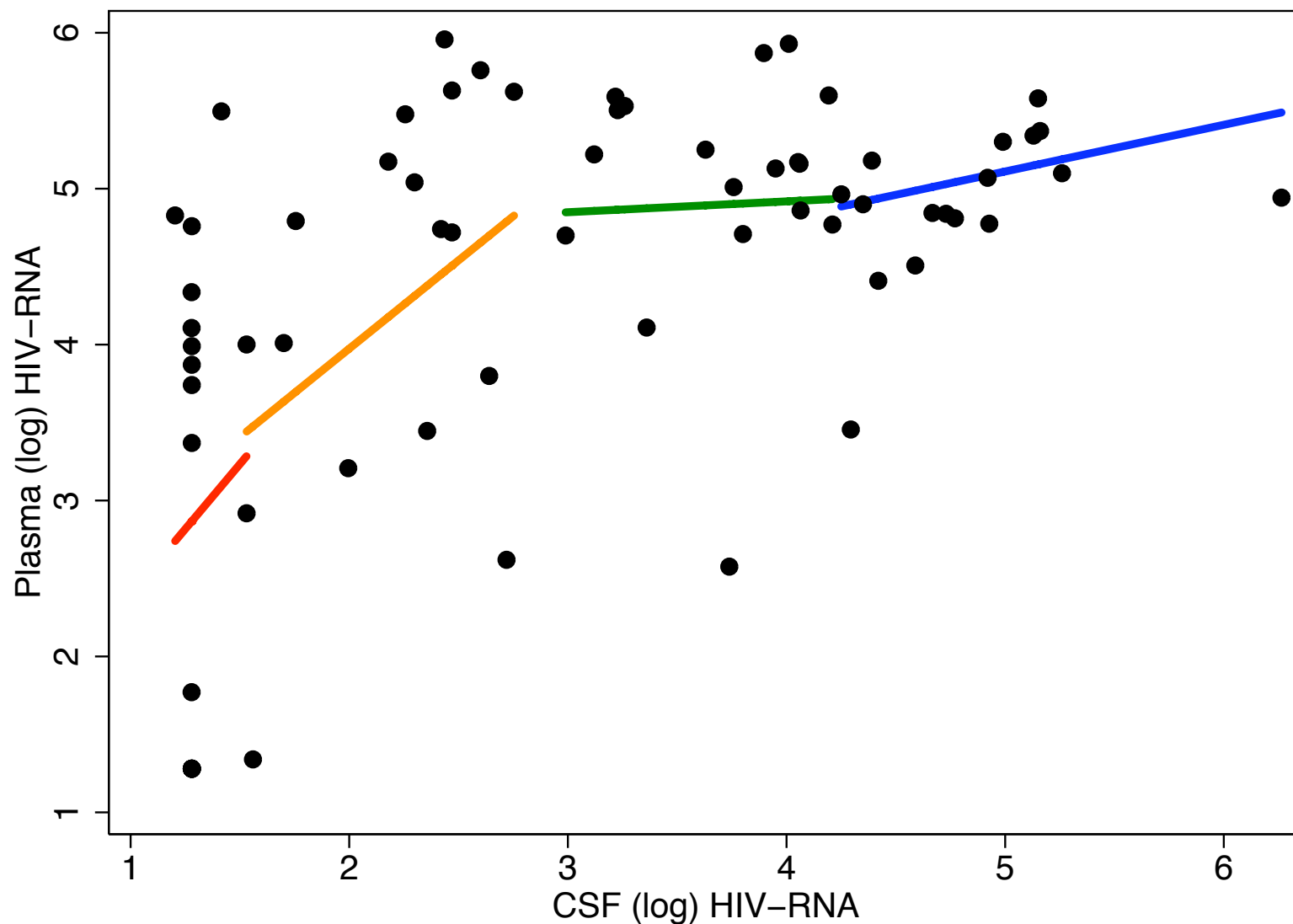
* Plot result (next slide)
. twoway (scatter hivrna csfrna , msize(vsmall)) (line hivrna2
csfrna, sort clpat(solid) lcolor(orange)),
plotregion(style(none)) ytitle("log Plasma HIV-RNA (Y)")
xtitle("log CSF HIV-RNA (X)") legend(order(1 2) label(1 "HIV-
RNA") label(2 "line") pos(6) ring(0)) saving(csfr3, replace)
```



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

Four local lines

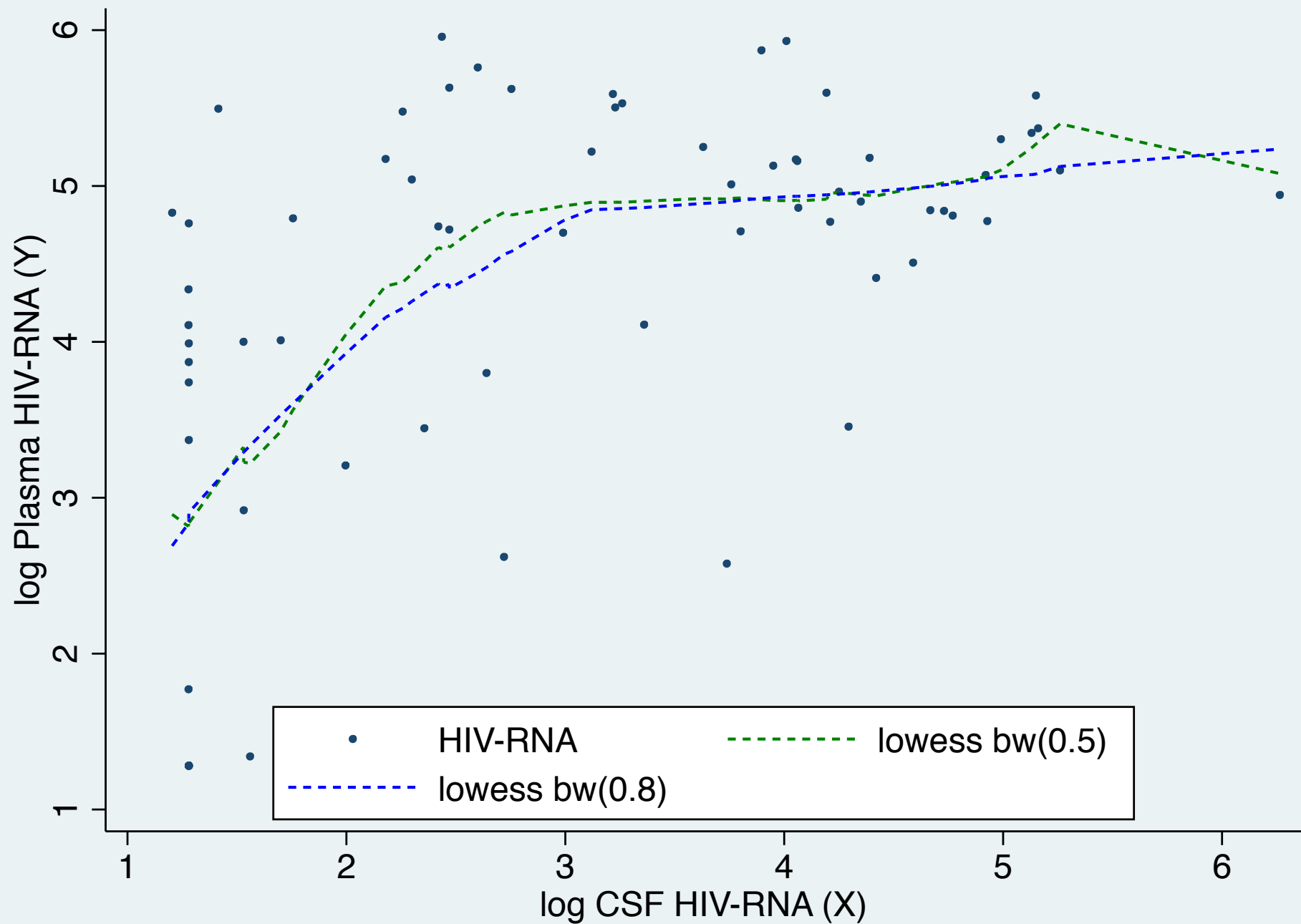


- Locally weighted *lowess* scatterplot smoothing joins estimates from local lines into a single smooth curve. The *bandwidth* parameter controls how much data is present in each “locale”

Lowess smoothing in Stata

- `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

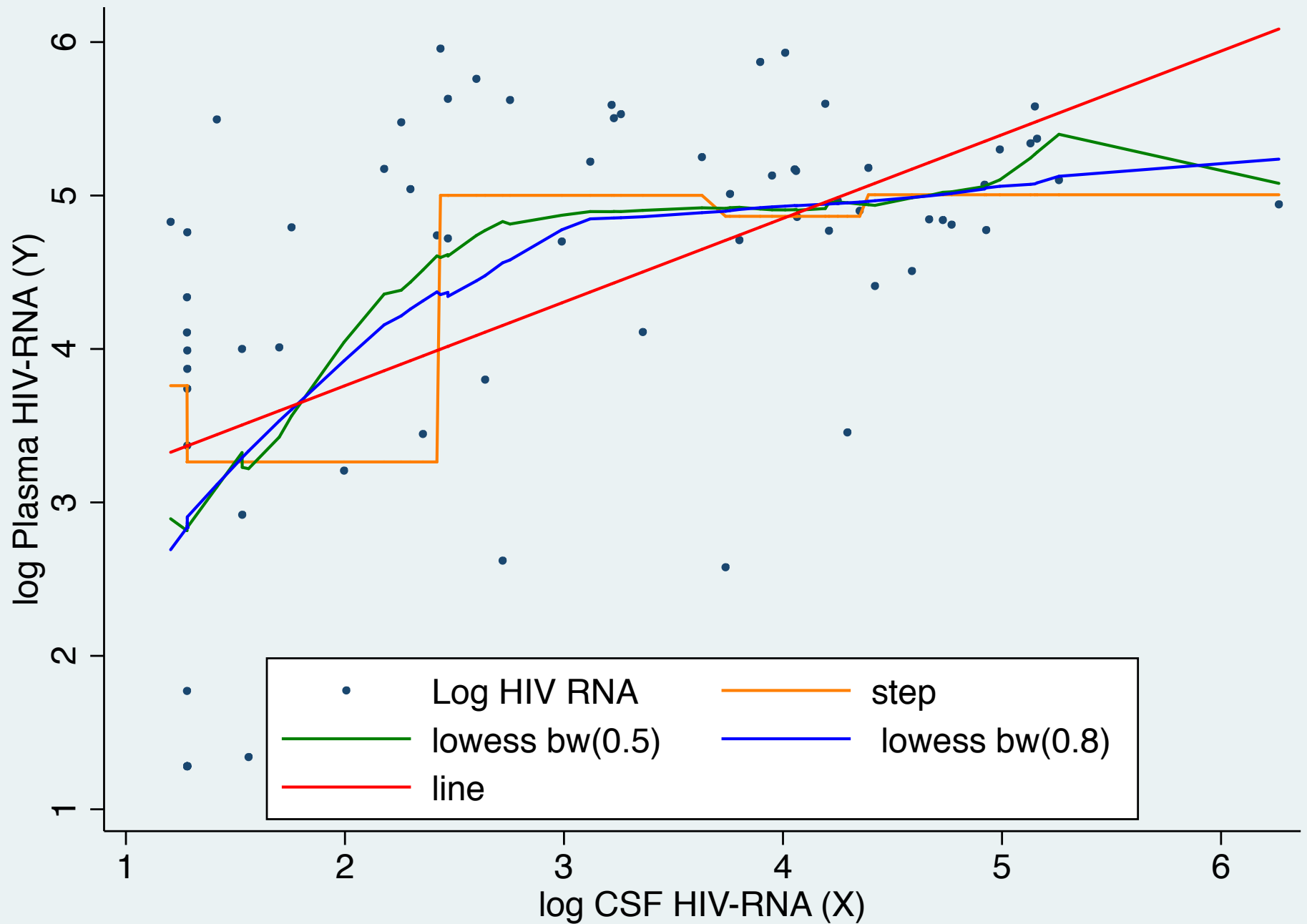
Lowess smooth for two bandwidths



How smooth?

- Smoothness in Lowess increases with bandwidth
- Too smooth: *imposes more or less linear fit, gives little information about non-linearity*
- Not smooth enough: *miss forest for the trees*
- Smooths may be unstable at plot margins “tail wagging”
- Can usually get by less smoothness in large samples, more in small ones

How smooth?



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:

$$E[y|x_1, x_2, \dots, x_p] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

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

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

- 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:

$$E[y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

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

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

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

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

- **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		
Model	1412.50418	1	1412.50418	F(1, 2030) = 14.97		
Residual	191605.195	2030	94.3867954	Prob > F = 0.0001		
Total	193017.699	2031	95.0357946	R-squared = 0.0073		
				Adj R-squared = 0.0068		
				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)

$$E[\text{glucose} | \text{exercise} = 0] = \beta_0 = 97.36104$$

- For exercisers (exercise = 1)

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

- 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		
Model	13153.784	2	6576.89199	F(2, 2027) = 74.14		
Residual	179802.433	2027	88.7037162	Prob > F = 0.0000		
Total	192956.217	2029	95.0991704	R-squared = 0.0682		
				Adj R-squared = 0.0673		
				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 -1.7)
- 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.

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

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

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

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

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

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

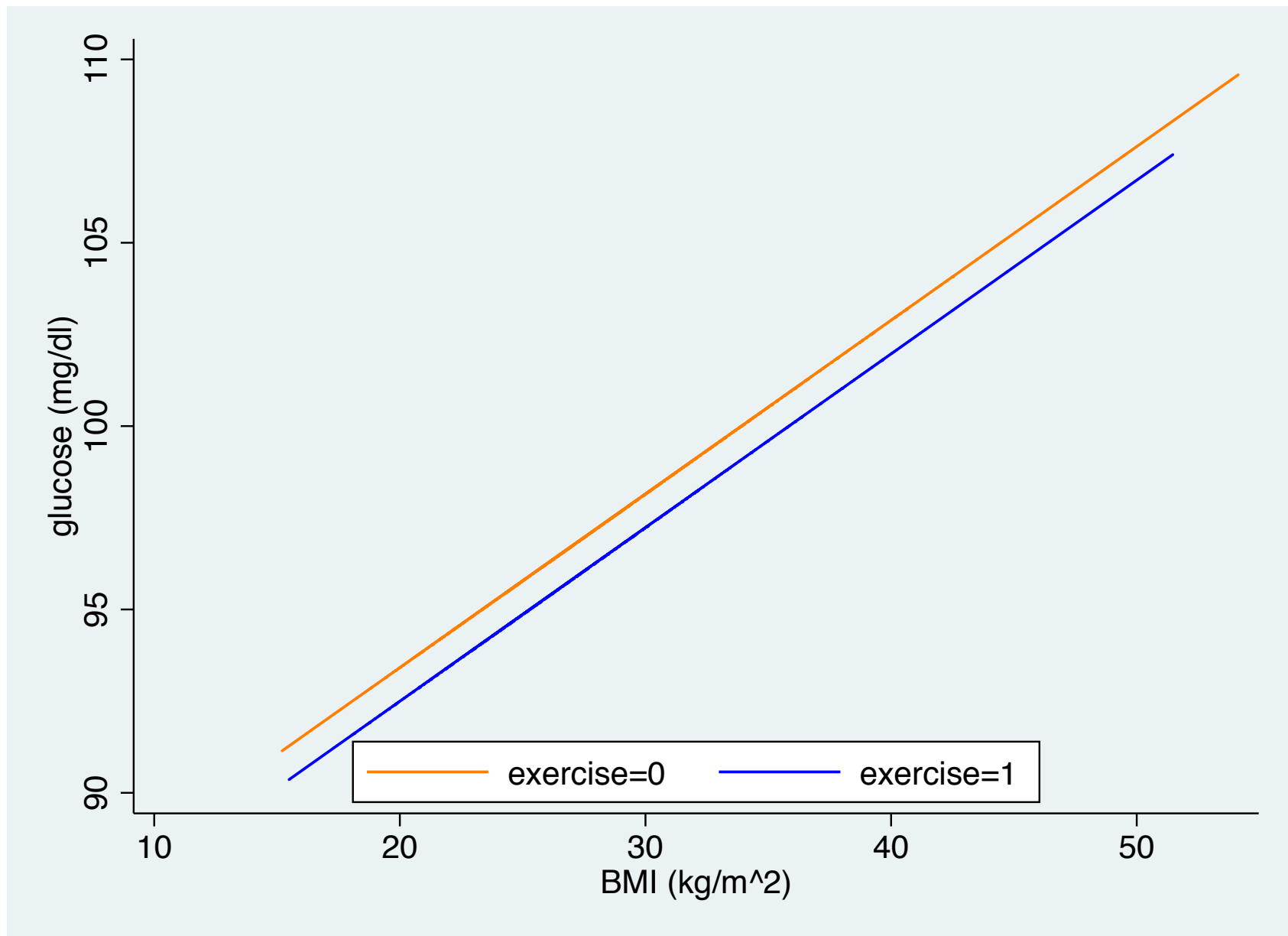
- Specifies a linear relationship between glucose and BMI for each level of exercise

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 I:** average glucose level related to BMI for two levels of exercise



Interpreting a Binary Predictor in a model with an additional continuous predictor (centered BMI):

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

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

glucose	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
cBMI	.4736147	.0411907	11.50	0.000	.3928342	.5543952
exercise	-.9172885	.4298059	-2.13	0.033	-1.760196	-.0743811
_cons	97.04504	.2745954	353.41	0.000	96.50653	97.58356

- Intercept coefficient gives average glucose (mg/dL) for individuals with exercise=0 and with average BMI (kg/m²)
- Coefficient for cBMI gives change in average glucose (mg/dL) for a unit increase in BMI (kg/m²), for either exercise group
- Coefficient for exercise gives difference in average glucose between those with exercise=1 and exercise=0, among individuals with a fixed BMI (e.g. the mean)

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.

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma_{y|\mathbf{x}}^2}{(n-1)\sigma_{x_j}^2(1-r_j^2)}$$

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

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

r_j^2 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 and 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

- **Example 2:** average glucose level related to BMI and HDL

. regress glucose BMI HDL if diabetes==0

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

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

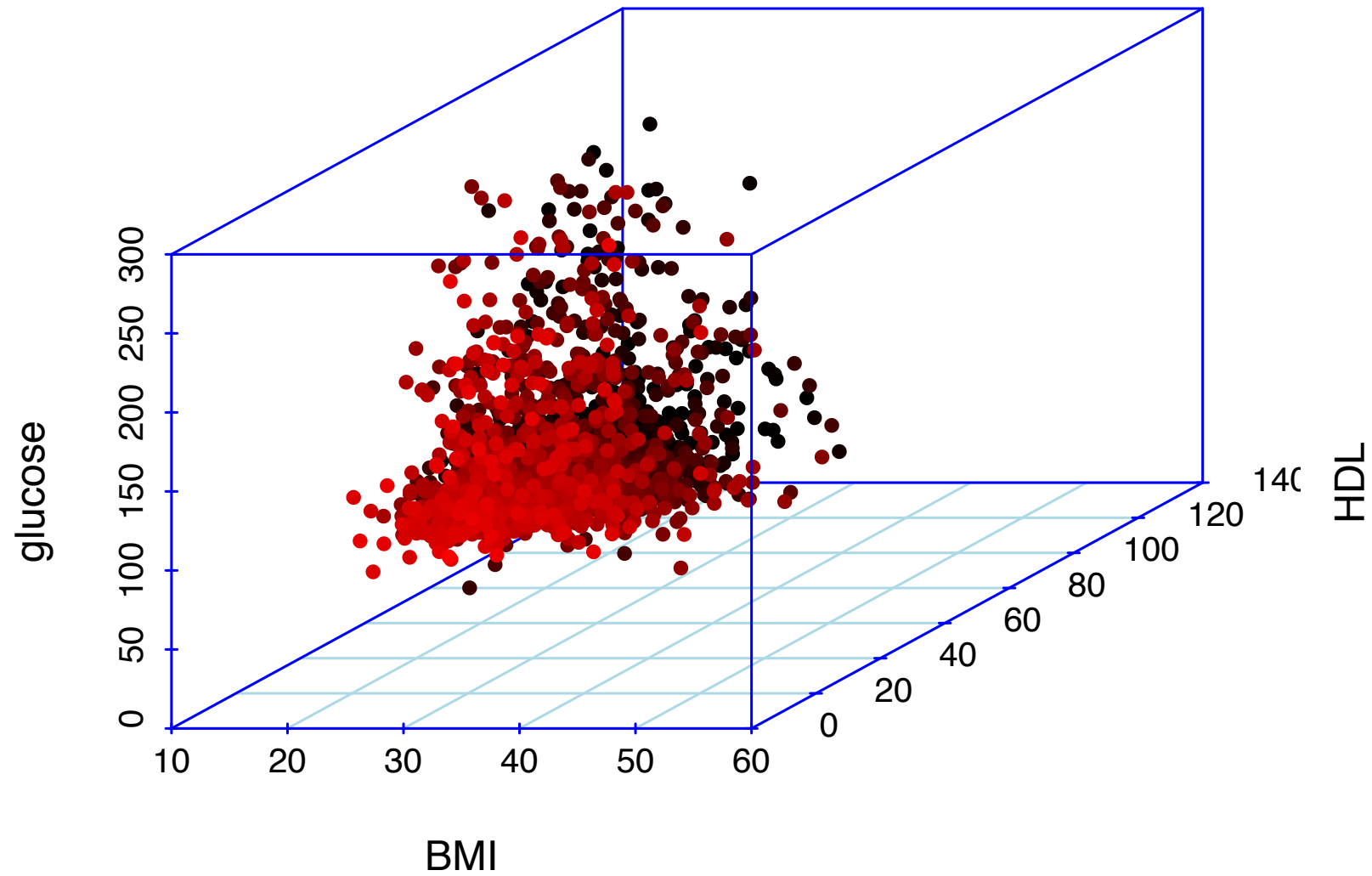
Specifies a linear relationship between glucose and HDL for any fixed value of BMI

- example: BMI=50:

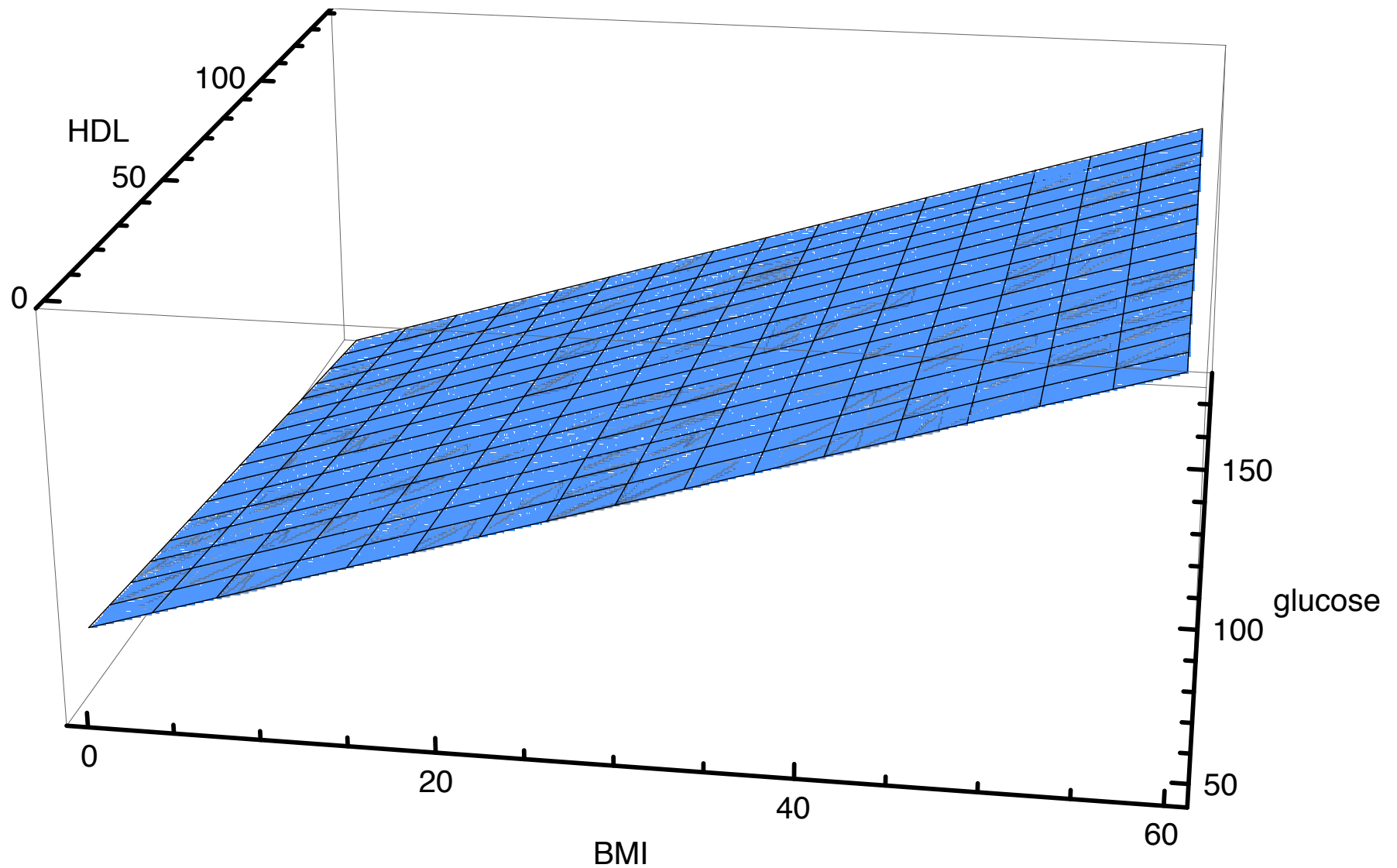
$$E[\text{glucose}|\text{BMI}=50, \text{HDL}] = \beta_0 + \beta_1 50 + \beta_2 \text{HDL} = 85.73492 + 0.468405 \times 50 - 0.0393889 \text{ HDL}$$

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
```

Coefficients for BMI and HDL unchanged
from previous model

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

Source	SS	df	MS
Model	13237.1205	2	6618.56023
Residual	179454.635	2020	88.8389284
Total	192691.756	2022	95.2976043

Number of obs = 2023
F(2, 2020) = 74.50
Prob > F = 0.0000
R-squared = 0.0687
Adj R-squared = 0.0678
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

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
```

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

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

Source	SS	df	MS
Model	13237.1203	2	6618.56017
Residual	179454.635	2020	88.8389284
Total	192691.756	2022	95.2976043

Number of obs = 2023
F(2, 2020) = 74.50
Prob > F = 0.0000
R-squared = 0.0687
Adj R-squared = 0.0678
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

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 = $10 \times \text{coef. for BMI}$, coef. for HDL_50 = $50 \times \text{coef. for HDL}$
- 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 of 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
Model	13237.1204	2	6618.56022
Residual	179454.635	2020	88.8389284
Total	192691.756	2022	95.2976043

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

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

Number of obs = 2023
F(2, 2020) = 74.50
Prob > F = 0.0000
R-squared = 0.0687
Adj R-squared = 0.0678
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

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	
Model	13237.1205	2	6618.56023	F(2, 2020) =	74.50
Residual	179454.635	2020	88.8389284	Prob > F =	0.0000
				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)

Use of `lincom` command to estimate outcome means for specified predictor values

- Example: conditional means for `exercise = 1` & `exercise = 0`, for `BMI=10`
- The `lincom` command is issued immediately after a regression model is fitted (slide 34)
- Gives the estimated mean outcome (and 95% CI) using a linear combination of coefficients

$$E[\text{glucose} | \text{exercise}=1, \text{BMI}=10] = \beta_0 + \beta_1 + \beta_2 10$$

$$E[\text{glucose} | \text{exercise}=0, \text{BMI}=10] = \beta_0 + \beta_2 10$$

```
. quietly regress glucose i.exercise BMI if diabetes==0
. lincom _cons + 1.exercise + BMI*10
```

```
( 1)  1.exercise + 10*BMI + _cons = 0
```

glucose	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
(1)	87.76106	.7605258	115.40	0.000	86.26957	89.25255

```
. lincom _cons + BMI*10
```

```
( 1)  10*BMI + _cons = 0
```

glucose	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
(1)	88.67835	.8037585	110.33	0.000	87.10207	90.25463

$$E[\text{glucose}|\text{exercise}=1, \text{BMI}=10] - E[\text{glucose}|\text{exercise}=0, \text{BMI}=10] = \beta_1$$

```
. lincom exercise
```

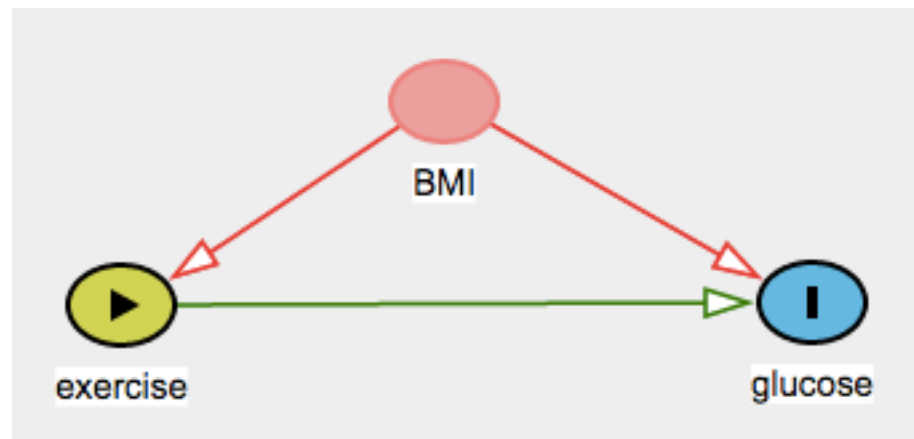
```
( 1)  exercise = 0
```

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

- The estimated regression coefficient $\beta_1 = -0.913$ estimates the conditional effect of exercise on glucose (expressed as a difference) *for any fixed BMI*

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

- 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:

$$E[\text{glucose}(\text{exercise}=1)] - E[\text{glucose}(\text{exercise}=0)]$$

- This difference is a *marginal causal effect* (with respect to BMI in this case - 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 of assignment to exercise).
- 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 previous model (adjusting for BMI) to estimate the marginal average glucose for the two values of exercise in the entire sample

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

```
. margins exercise
```

```
Predictive margins
```

```
Number of obs   =       2030
```

```
Model VCE       : OLS
```

```
Expression      : Linear prediction, predict()
```

		Delta-method		t	P> t	[95% Conf. Interval]	
		Margin	Std. Err.				
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: Using regression to estimate the marginal causal effect of exercise on glucose

- The difference between the marginal means give the average causal effect:

```
. margins, dydx(exercise)
```

Average marginal effects

Model VCE : OLS

Number of obs = 2030

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 model
- For linear models, the marginal causal effect can be *estimated directly* by a regression coefficient (not necessarily the case for nonlinear models such as logistic regression)
- So, regression coefficients from linear models can be interpreted as both conditional and marginal (depending on context)

Example: Binary Predictors with Reversed Categories

- Define a new exercise variable as follows:
 - 1 : exercise < 3 times / week ; 0 : exercise $\geq$ 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
Model	13153.784	2	6576.89199	F(2, 2027)	=	74.14
Residual	179802.433	2027	88.7037162	Prob > F	=	0.0000
				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		
Model	13153.784	2	6576.89199	F(2, 2027) = 74.14		
Residual	179802.433	2027	88.7037162	Prob > F = 0.0000		
Total	192956.217	2029	95.0991704	R-squared = 0.0682		
				Adj R-squared = 0.0673		
				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