

# Bivariate Relationships: Exploratory Methods and Simple Linear Regression

January 12, 2010  
Biostatistics 208

# This week:

- Announcements (Stata 11 vs 10)
- Exploring bivariate relationships
- Introduction to linear regression
- Checking the single-predictor linear model using exploratory methods

# Why regression?

- In RCTs, we infer a causal effect from differences in means between randomized groups
- Regression summarizes how mean of outcome depends on predictors
  - *multiple regression does this accounting for effects of confounders*
- For a binary or categorical predictor, the summary in a simple linear regression is equivalent to the comparison between means in the groups, *as in the t-test or ANOVA*

# Example: sample ( $n=221$ ) from the HERS study

## Data description from Stata:

```
. describe
```

```
Contains data from ~/Documents/teaching/c2010/atcr/lecture2/data/hers-sample.dta
```

```
obs:          221
```

```
vars:          3
```

```
18 Jan 2005 08:36
```

```
size:         2,431 (99.9% of memory free)
```

```
-----
```

variable name	storage type	display format	value label	variable label
<b>smoking</b>	byte	%9.0g	noyes	current smoker
<b>waist</b>	float	%9.0g		waist circumference (cm)
<b>hdl</b>	int	%9.0g		hdl cholesterol (mg/dl)

```
-----
```

```
Sorted by:
```

*(continued on next slide)*

# 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  
                                  5

# 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 example:

- Relationship between HDL cholesterol and a single binary predictor representing smoking status (Y/N)
- Relationship between HDL cholesterol and a continuous measure of waist circumference (cm)

# Linear regression

- Mean of outcome assumed to change linearly with a continuous predictor
- Uses all the data to estimate mean at any point; *borrow strength across points*
- Predictor can be of any type: *binary, categorical, a count, as well as continuous*
- Normality of continuous *predictor* not required
- Linearity assumption can be worked around via transformations

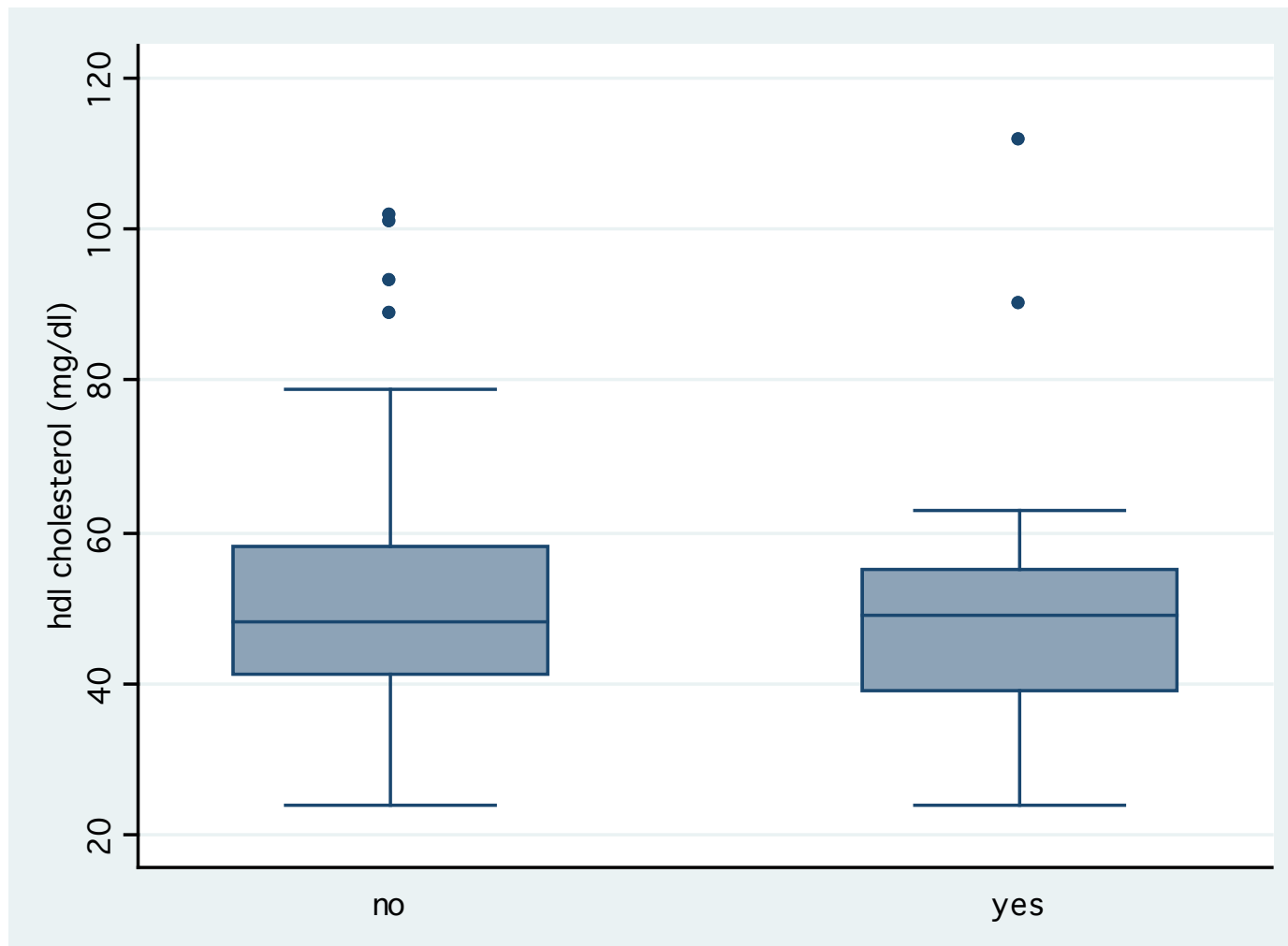
# Systematic part of linear model

- Model represents the mean of  $y$  as a function of  $x$ 
  - $E(y|x) = \beta_0 + \beta_1 x$
- Values of  $E(y|x)$  lie along the regression line
- Dependence on predictor summarized by the slope of the line,  $\beta_1$ 
  - $\beta_1$  measures change in outcome for a unit increase in  $x$
- Intercept  $\beta_0$  gives mean of  $y$  when  $x = 0$

# Random part of linear model

- $Y = \beta_0 + \beta_1 X + \varepsilon$ 
  - Data = conditional mean + random error
- Assumptions about random errors ( $\varepsilon$ )
  - independent
  - mean zero at all values of  $x$
  - equal variance at all values of  $x$
  - normally distributed (at least in small samples)
- Assumptions required for valid inference (i.e.  $p$ -values & confidence intervals)

# HDL by smoking status



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

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	<b>50.26943</b>	.9490756	13.18498	48.39748	52.14138
yes	26	<b>50.34615</b>	3.578114	18.24487	42.97689	57.71542
combined	219	50.27854	.9345828	13.83056	48.43656	52.12051
diff		<b>-.0767238</b>	2.895971		<b>-5.784556</b>	<b>5.631109</b>

**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:

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	<b>50.26943</b>	.9490756	13.18498	48.39748	52.14138
yes	26	<b>50.34615</b>	3.578114	18.24487	42.97689	57.71542
combined	219	50.27854	.9345828	13.83056	48.43656	52.12051
diff		<b>-.0767238</b>	2.895971		<b>-5.784556</b>	<b>5.631109</b>

**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:

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	<b>50.26943</b>	.9490756	13.18498	48.39748	52.14138
yes	26	<b>50.34615</b>	3.578114	18.24487	42.97689	57.71542
combined	219	50.27854	.9345828	13.83056	48.43656	52.12051
diff		<b>-.0767238</b>	2.895971		<b>-5.784556</b>	<b>5.631109</b>

**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:

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	<b>50.26943</b>	.9490756	13.18498	48.39748	52.14138
yes	26	<b>50.34615</b>	3.578114	18.24487	42.97689	57.71542
combined	219	50.27854	.9345828	13.83056	48.43656	52.12051
diff		<b>-.0767238</b>	2.895971		<b>-5.784556</b>	<b>5.631109</b>

**diff = mean(no) - mean(yes)**

t = -0.0265

Ho: diff = 0

degrees of freedom = 217

Ha: diff < 0

Pr(T < t) = 0.4894

Ha: diff != 0

**Pr(|T| > |t|) = 0.9789**

Ha: diff > 0

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:

# 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 two-level categorical predictor not coded 0/1 should be recoded as such before entering in a regression model
- 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

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

```
. regress hdl smoking
```

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

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

# Note on post-estimation commands in Stata:

When we fit regression models, frequently we want to use the results to perform additional tests or make further estimates. Stata handles these tasks via *postestimation* commands.

**Example:** using `lincom` to estimate the mean outcome value for smokers and non-smokers in the previous example.

Issued immediately after a `regress` command, `lincom` provides estimates of **linear combinations** 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)	<b>50.34615</b>	2.718635	18.52	0.000	<b>44.98784</b> <b>55.70446</b>

```
. lincom _cons
```

hdl	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
(1)	<b>50.26943</b>	.9978353	50.38	0.000	<b>48.30274</b> <b>52.23612</b>

Note : the second command returns the intercept 15

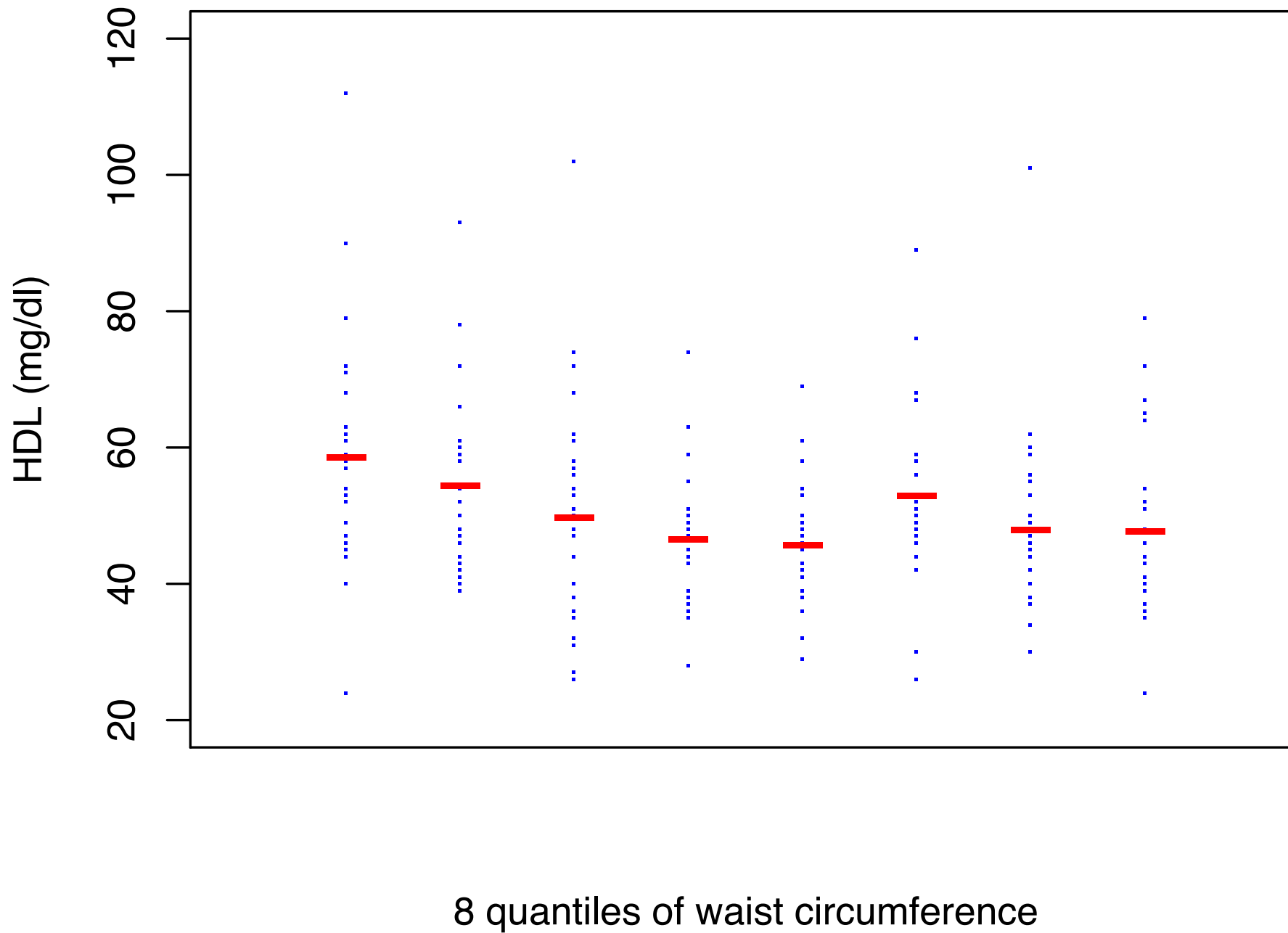
# Analysis issue:

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

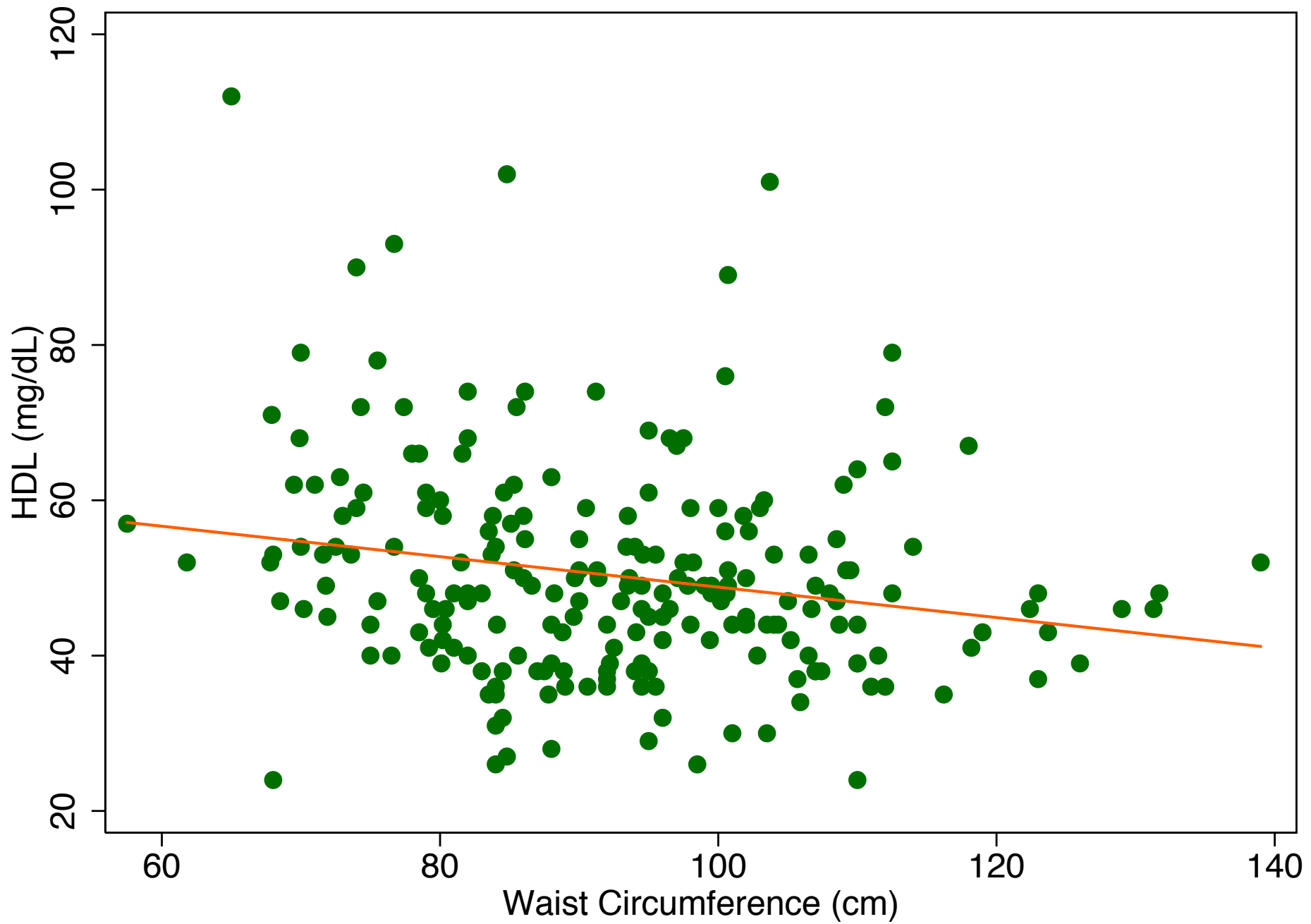
# Simple linear regression with a continuous predictor

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

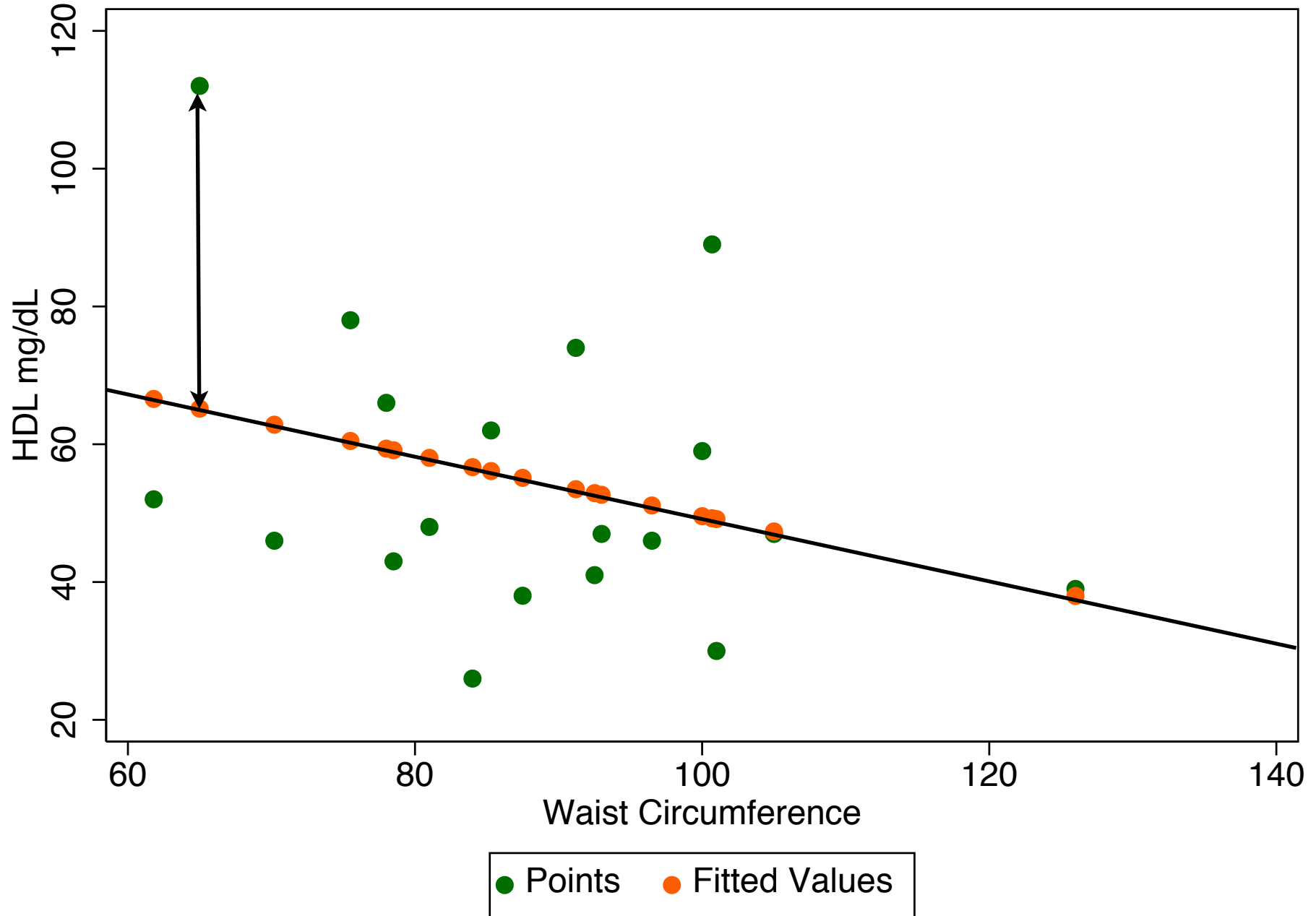
# Mean HDL by octile of waist circumference



# Linear model: line of means



# Data point, 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)* (a residual is illustrated by the vertical line in the previous slide)
- Efficient - *minimum variance in class of estimators*
- Maximum likelihood estimates (MLEs) if outcome is normal
- Drawback of OLS: sensitive to outliers

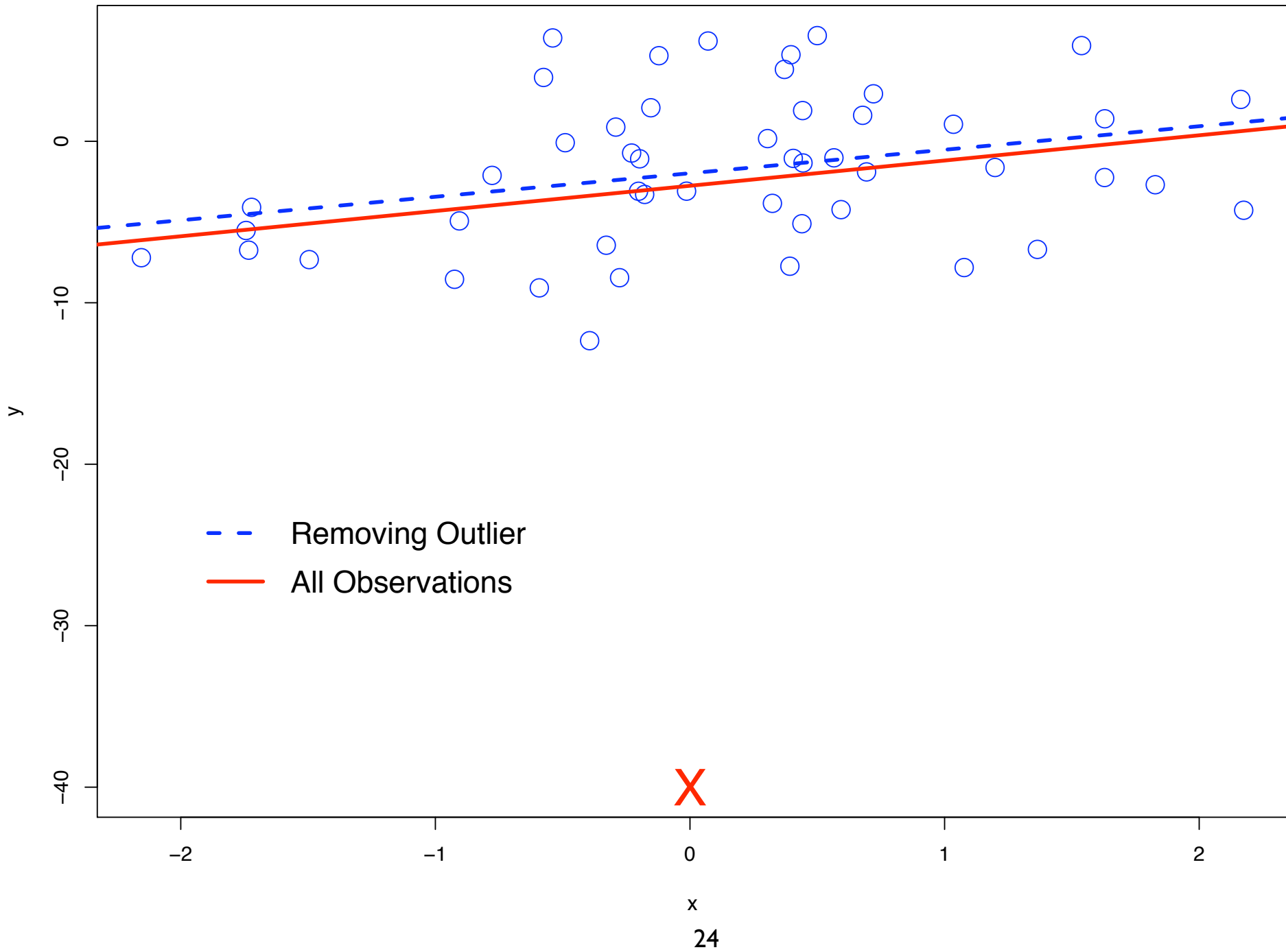
# Outliers and influential points

- Influential points unduly affect the slope estimate
- 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

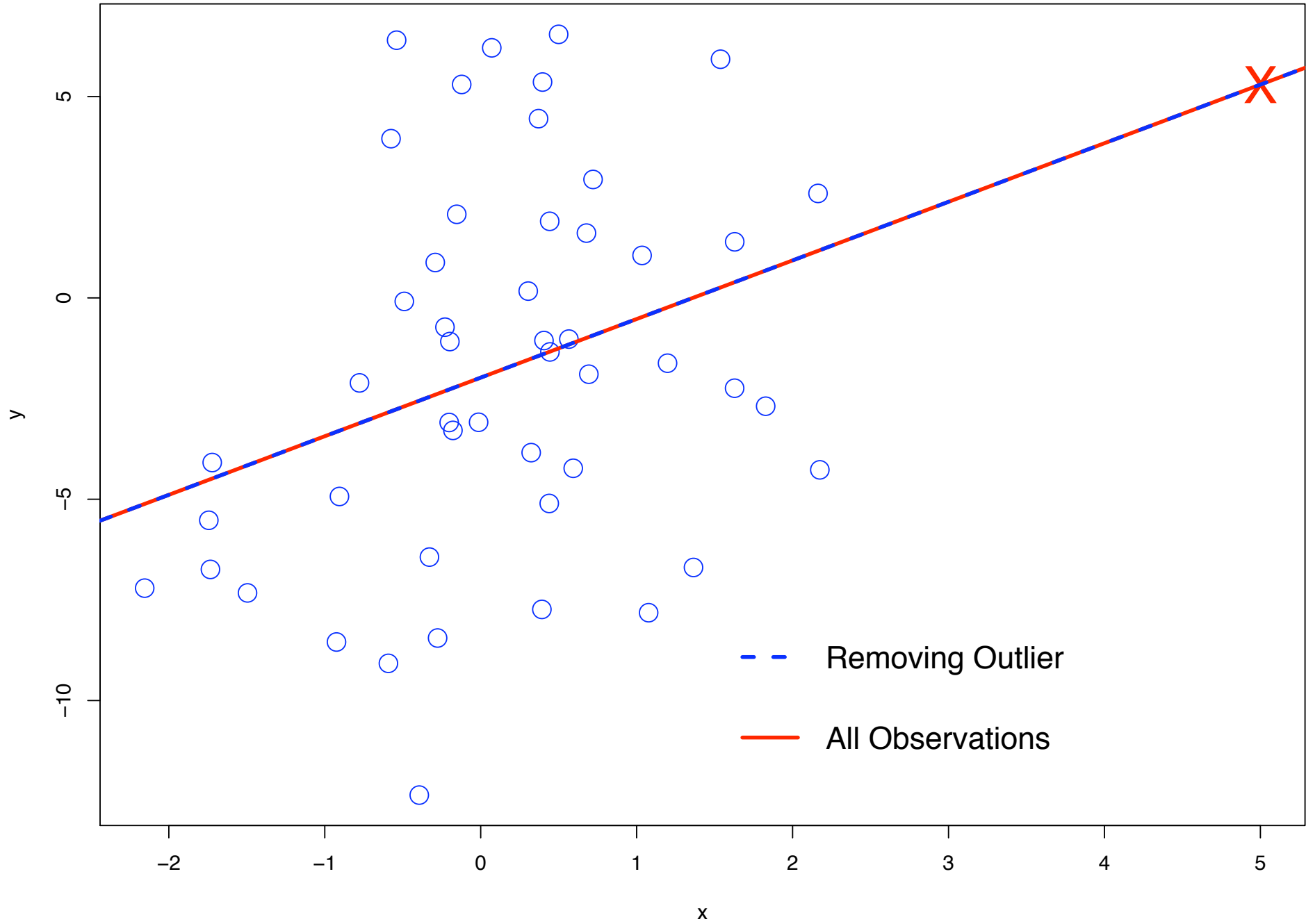
# Scatterplots as model diagnostics

- Most useful when both outcome and predictor are continuous
  - Detect outliers
  - Check linearity assumption, determine need for transformation

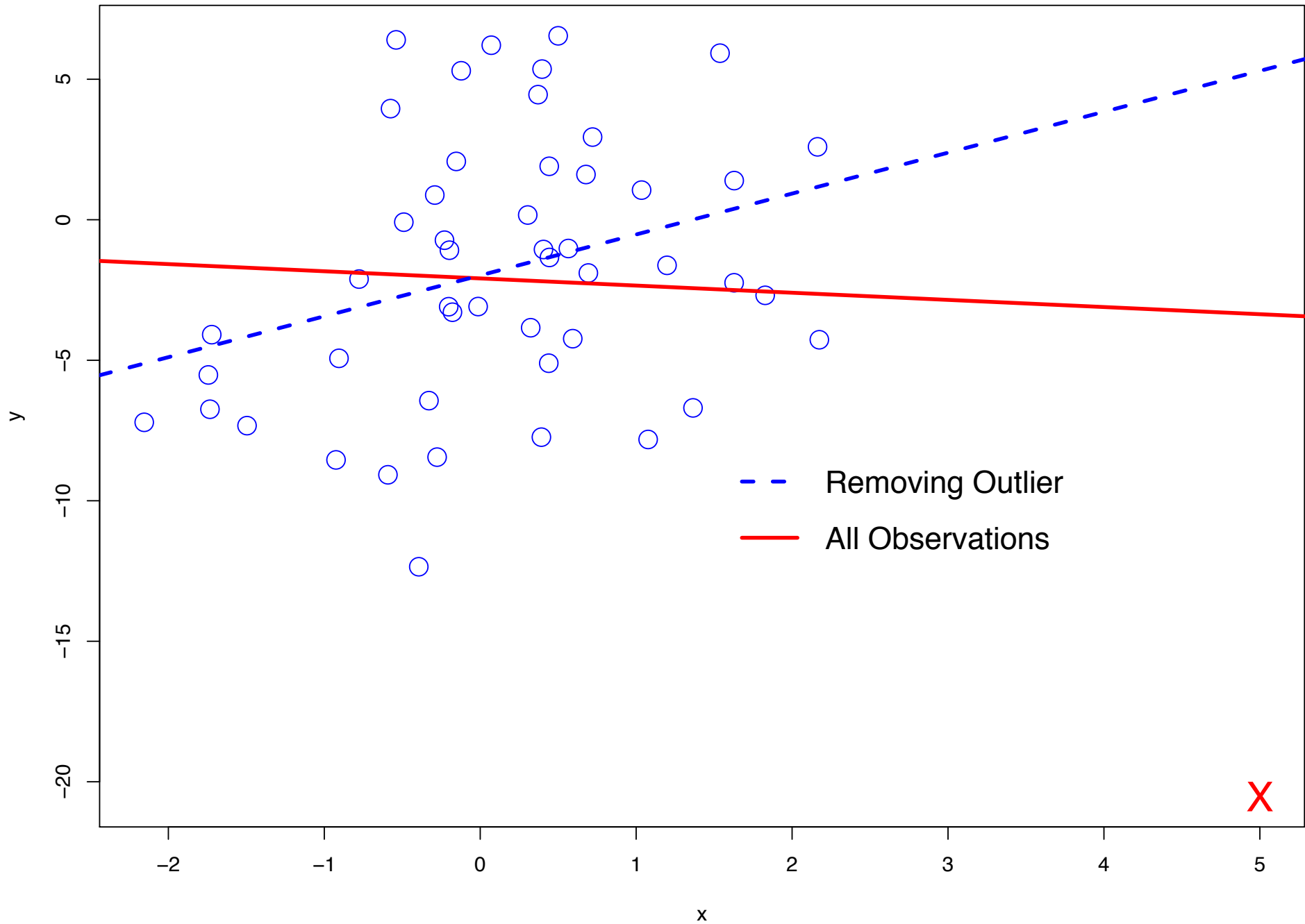
# Benign outlier



# High leverage point



# Influential point



# 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
- 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 about

# Questions a regression can answer

1. How does mean of outcome depend on predictor?
2. How precisely is this dependence estimated, and is the estimated association statistically significant?
3. How much of variation is explained by variation in predictor?

# Dependence of HDL on waist circumference

. regress hdl waist

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

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

# Interpreting coefficients from the linear model

- Model represents the mean of HDL as a function of waist circumference
  - $\text{mean}(HDL \mid \text{waist circ.}) = 68.43 - 0.196 \times (\text{waist circ.})$
- 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.
  - $\text{mean}(HDL \mid \text{waist circ.} = 100) - \text{mean}(HDL \mid \text{waist circ.} = 99) =$   
 $[68.43 - 0.196 \times (100)] - [68.43 - 0.196 \times (99)] = -0.196$
- Use model to compute change in aver. HDL assoc. with a 10 unit increase in waist circ.
  - $\text{mean}(HDL \mid \text{waist circ.} = 100) - \text{mean}(HDL \mid \text{waist circ.} = 90) =$   
 $[68.43 - 0.196 \times (100)] - [68.43 - 0.196 \times (90)] = -0.196 \times 10$   
 $= -1.96$

# Precision and statistical significance of slope estimate

. regress hdl waist

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

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

# Proportion of variation explained

```
. regress hdl waist
```

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

# Marginal and residual variance estimates

. regress hdl waist

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

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

# 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

# Centering to make intercept meaningful

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

# Model with observed waist circumference

```
. regress hdl waist
```

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

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

# Model with centered waist circumference

```
. egen meanwaist = mean(waist)
```

```
. gen cwaist = waist - meanwaist
```

```
. reg hdl cwaist
```

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

hdl	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
<b>cwaist</b>	<b>-.1960017</b>	.0630536	-3.11	0.002	-.3202776	-.0717258
<b>_cons</b>	<b>50.28492</b>	.9165519	54.86	0.000	48.47844	52.0914

# 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 centered and scaled (AKA standardized)
- Centering, and scaling by the SD of predictor leads to “standardized” regression coefficients

# Model with centered and scaled waist circumference

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

```
. reg hdl cswaist
```

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

hdl	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
<b>cswaist</b>	<b>-1.960017</b>	.6305362	-3.11	0.002	-3.202776	-.7172577
<b>_cons</b>	<b>50.28492</b>	.9165519	54.86	0.000	48.47844	52.0914

- HDL decreases by ~2 mg/dl for every 10 cm increase in waist size

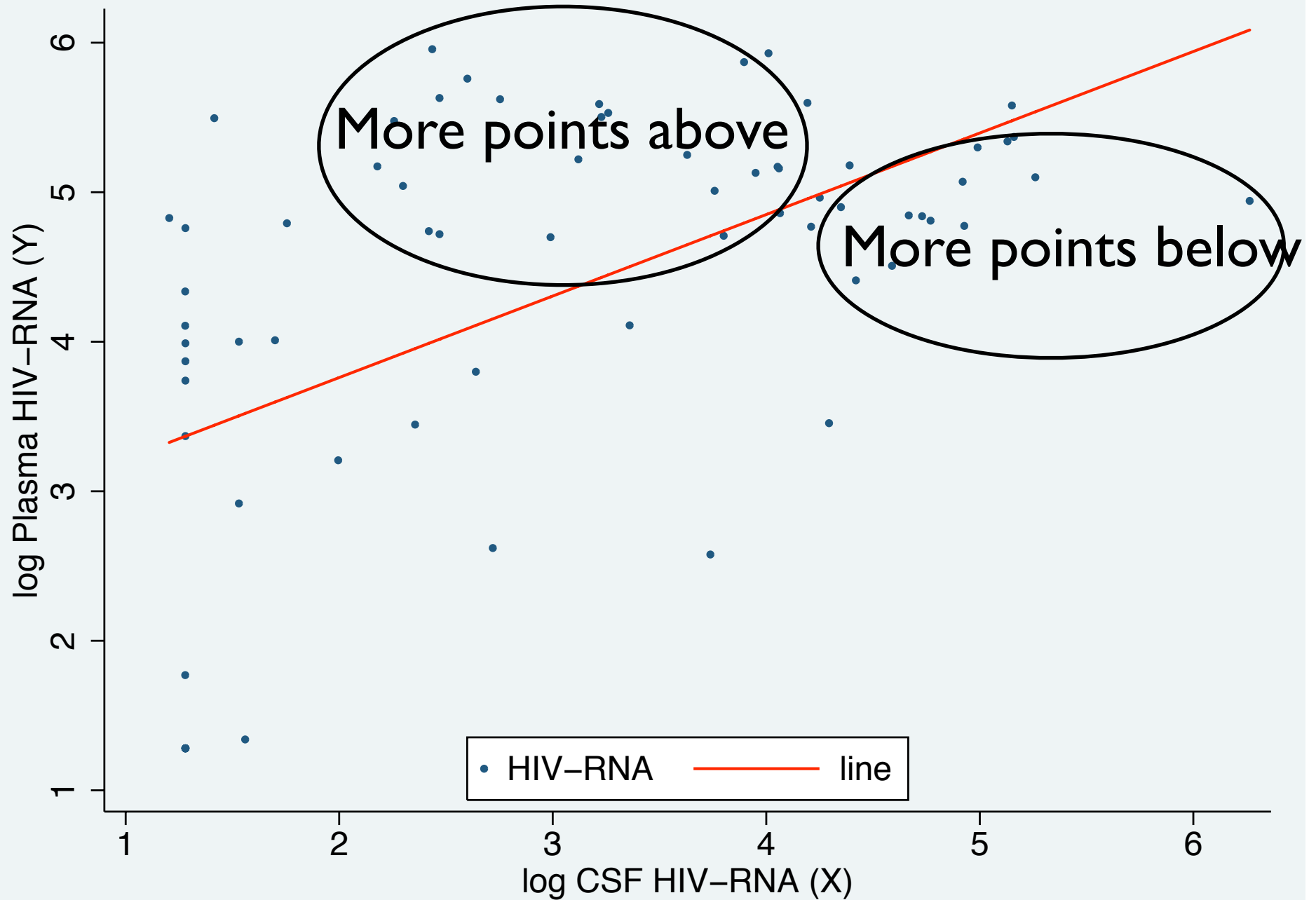
# Exploratory methods for assessing linearity

- 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 (lecture 6)

# Example: relationship between HIV viral load in plasma and CSF



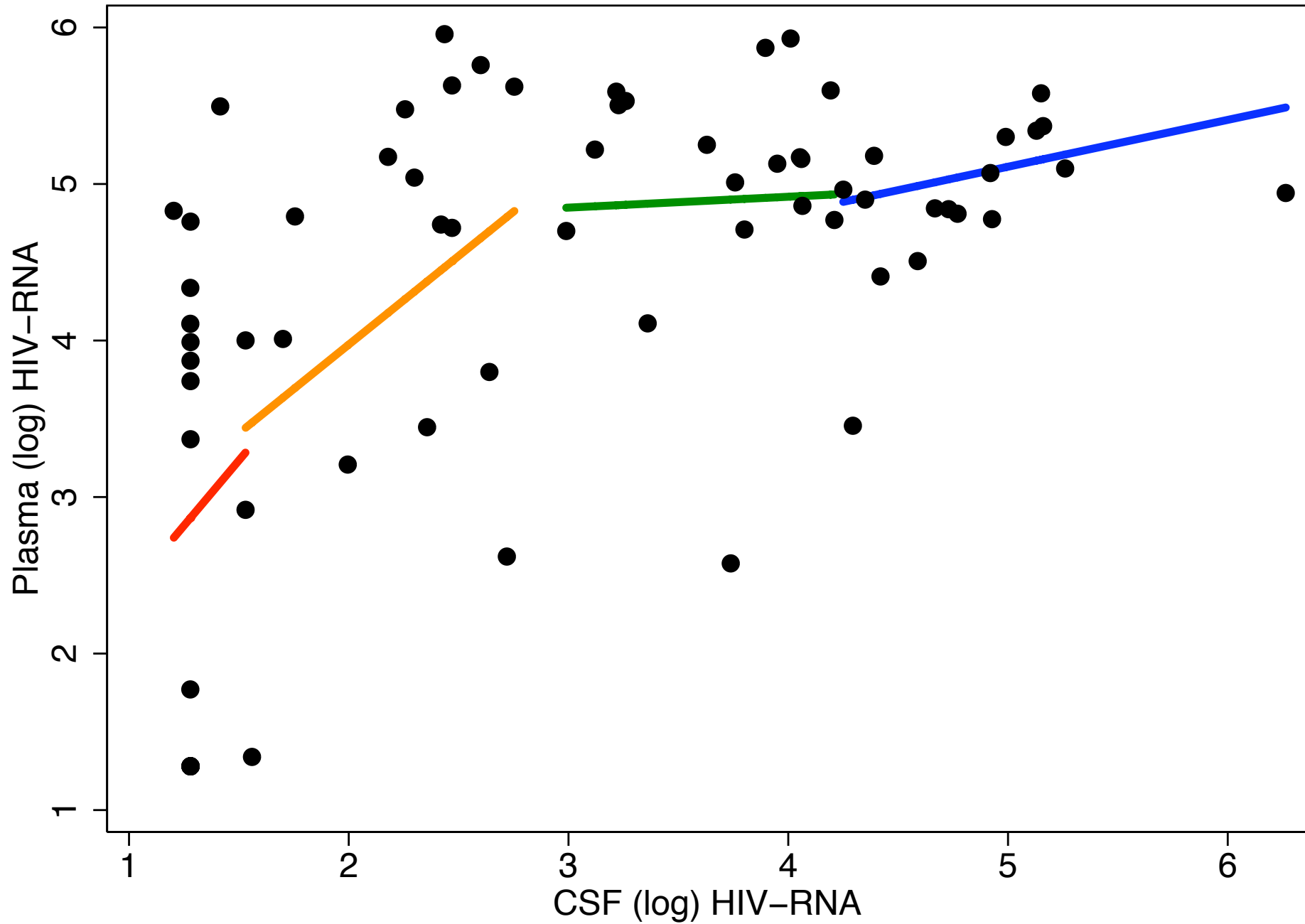
# Is a straight line adequate?



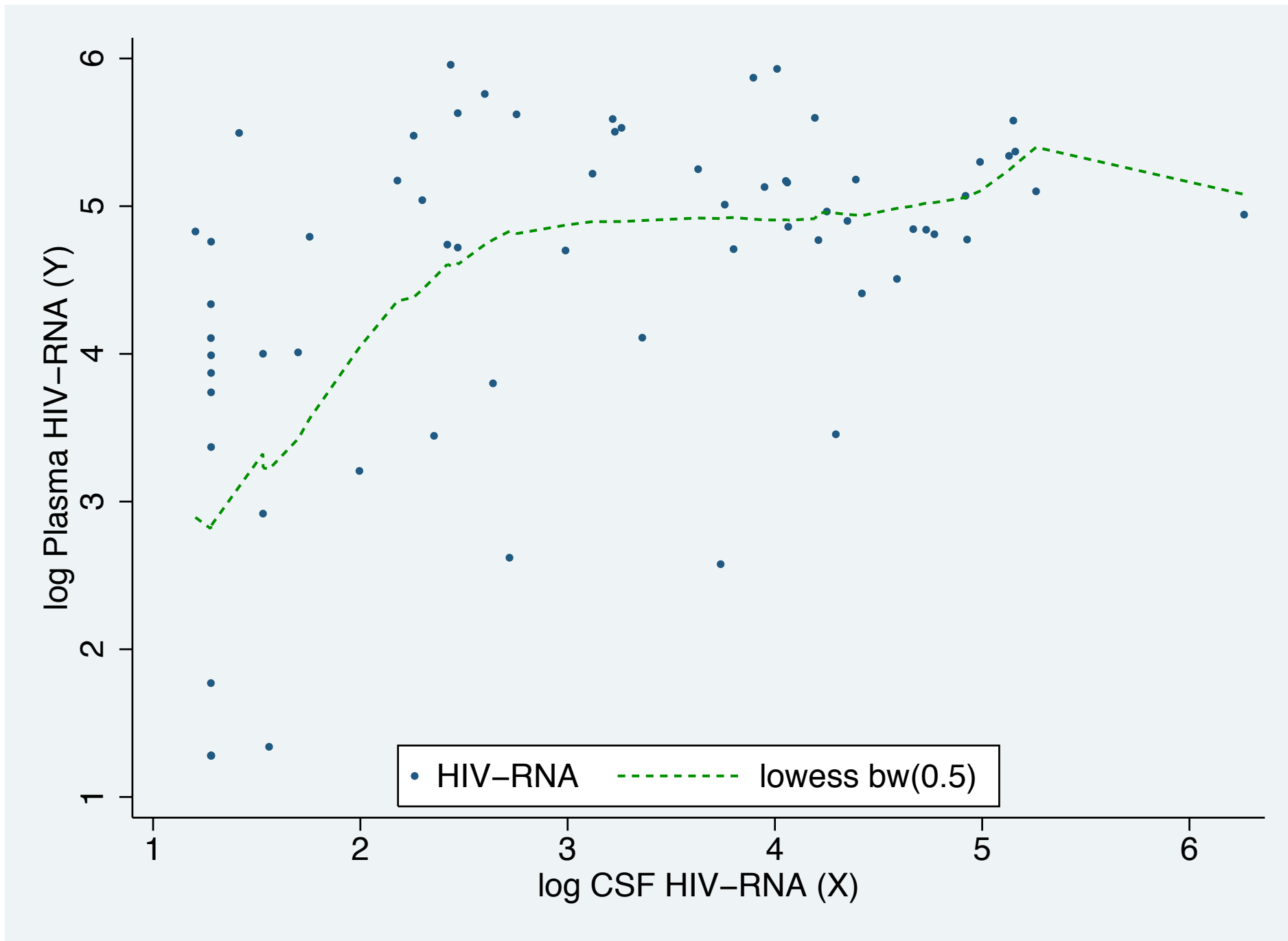
# Scatterplot smoother

- 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



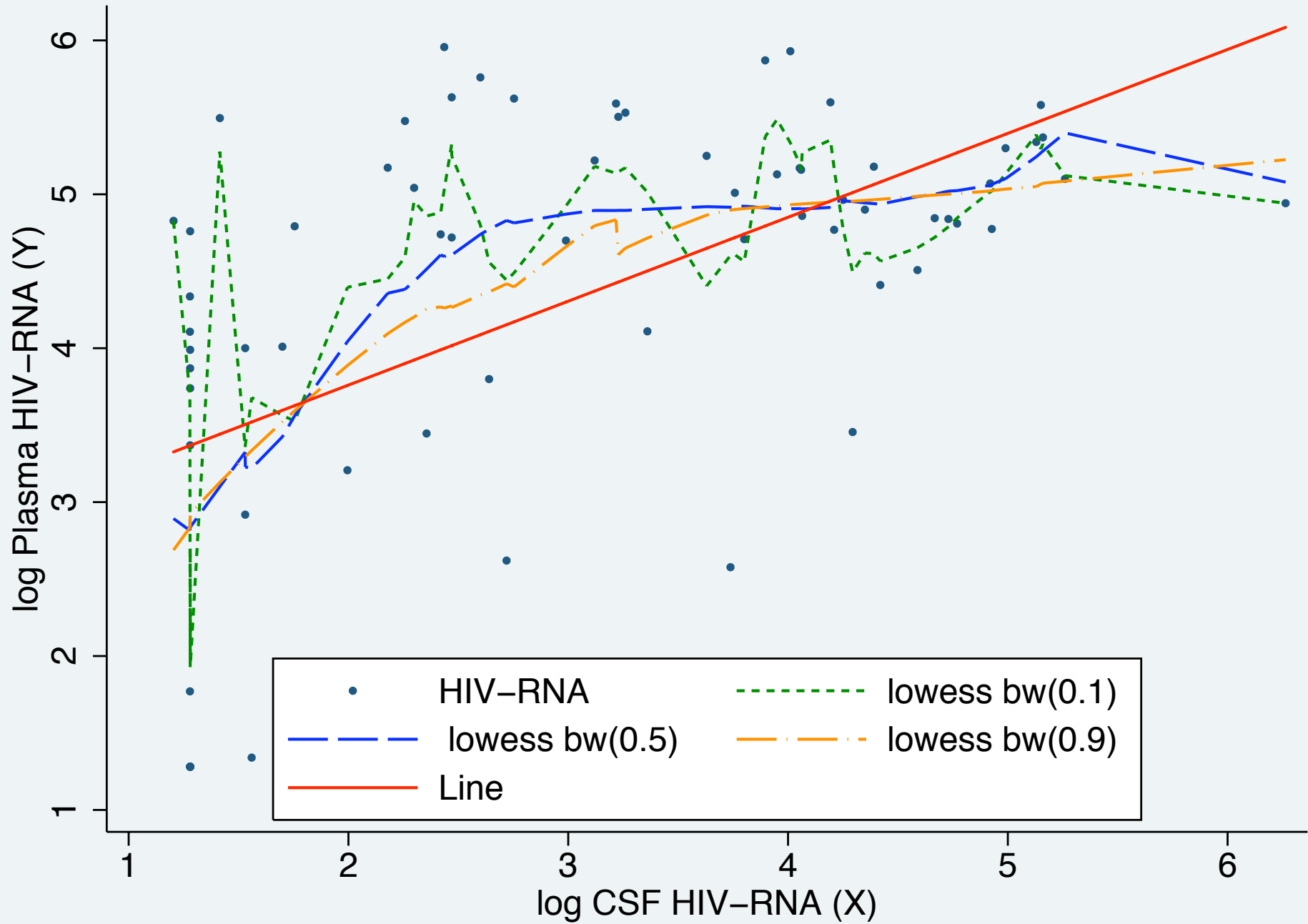
# Lowess smooth



# How smooth?

- Smoothness in Lowess increases with “bandwidth” (range 0 to 1)
- Too smooth: *imposes more or less linear fit, gives little information about non-linearity*
- Not smooth enough: *miss forest for the trees*
- Smooths commonly wag their tails
- Can use less smoothness in large samples, more in small ones

# How smooth?



# Stata commands

- `lowess csfrna hivrna, bw(0.5)`
- above command specifies a scatterplot with lowess smooth, specifying less smoothness than default (0.8)
- Using menus: *Graphics, Twoway Graphs*
- More about this in lecture 6

# Summary

- Bivariate analyses with continuous outcomes
- Exploratory bivariate techniques
  - *side-by-side boxplots*
  - *scatterplots with Lowess smooths*
- Simple linear regression
  - *rests on linearity and other assumptions*
  - *sensitive to outliers*
  - *provides CIs, p-values (unlike Lowess)*
- Exploratory methods also valuable for checking linearity assumption and for outliers