Stat 306:
Finding Relationships in Data.
Lecture 16
Sections 4.3

Chapter 4 – Variable selection and additional diagnostics

4.1 Variable Selection algorithms

4.2 Cross-validation and out-of sample assessment

4.3 Additional diagnostics

4.4 Transforms and nonlinearity

4.5 Diagnostics for data collected sequentially in time

In today's lecture:

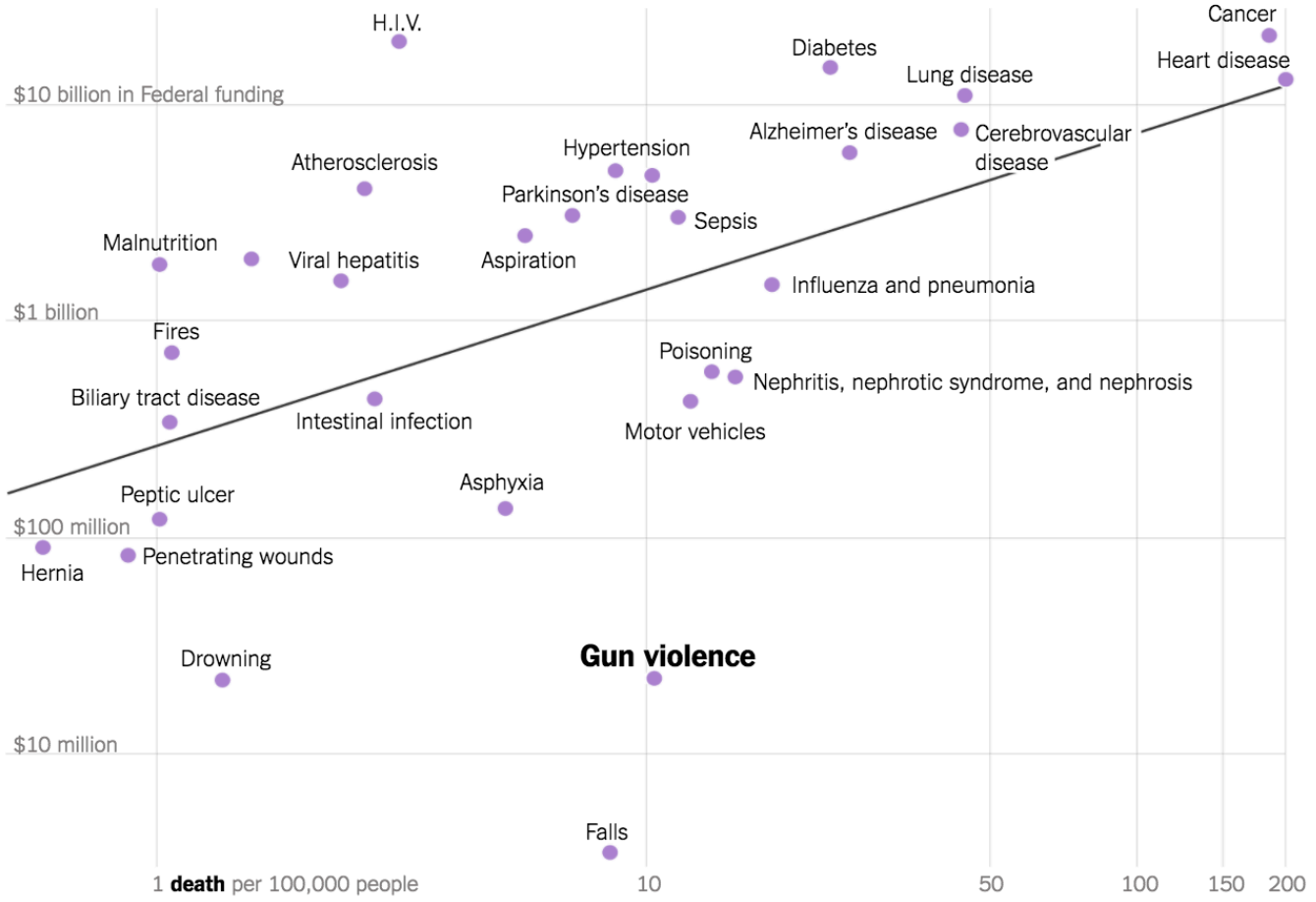
- **Leave-one-out cross validation from last lecture**
- **Start with a question about residuals.**
- **Abruptly change topics to discuss "Influence"**
- **Change topics again to talk about the "Hat" matrix**
- **Go back to Influence with the Hat matrix**
- **Then … I tell you the truth about residuals!**
- **Now that you know the truth, what should you do?**
- **Answer our simple question about residuals from the beginning.**
- **How does this help us understand NYT article on researhc for gun violence?**
- **BONUS: all this was useful for Leave-one-out cross validation**

The New York Times

There's an Awful Lot We Still Don't Know About Guns

By **QUOCTRUNG BUI** and **MARGOT SANGER-KATZ** MARCH 2, 2018

Federal funding for research on leading causes of death



Source: From 2004 to 2014, David Stark and Nigam Shah, Funding and Publication of Research on Gun Violence and Other Leading Causes of Death

For each model, we do 5-fold CV:

Metric:

Mean Absolute Prediction Error:

K-averaged metric = 40/5 = 8

Source: http://blog.goldenhelix.com/goldenadmin/cross-validation-for-genomic-prediction-in-svs/

Goal is

Prediction

4.2 Leave-one-out

Leave-one-out

Sample of size n, $(y_i, x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$.

For $i = 1, \ldots, n$, delete the ith observation (ith row of data set) and fit a regression with $n-1$ observations/cases.

Let the least squares regression vector be denoted as $\hat{\boldsymbol{\beta}}_{-i}$. Let $\mathbf{x}_i^T = (1, x_{i1}, \ldots, x_{ip})$ be ith row of the $n \times (p+1)$ matrix $\mathbf{X}$. The prediction of the ith response based on the remaining observations is $\hat{y}_{i|-i} = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{-i}$, and the prediction error is $y_i - \hat{y}_{i|-i}$.

The cross-validated root mean square (prediction) error is:

$$CVRMSE_{leaveoneout}(x_1, \ldots, x_p) = \sqrt{n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_{i|-i})^2}\,.$$

$CVRMSE_{leaveoneout}(\mathbf{x}_J)$ can be defined for different subsets of the explanatory variables. [$\mathbf{x}_J = (x_j : j \in J)$= variables indexed by set J]

What is an outlier?

What is an outlier?

How big is a “big residual”?

What is an outlier?

How big is a “big residual”?

$$\epsilon_i \sim Normal(0, \sigma^2)$$

So maybe we could say that a big residual is…

$$\hat{\epsilon}_i > 2\hat{\sigma}$$

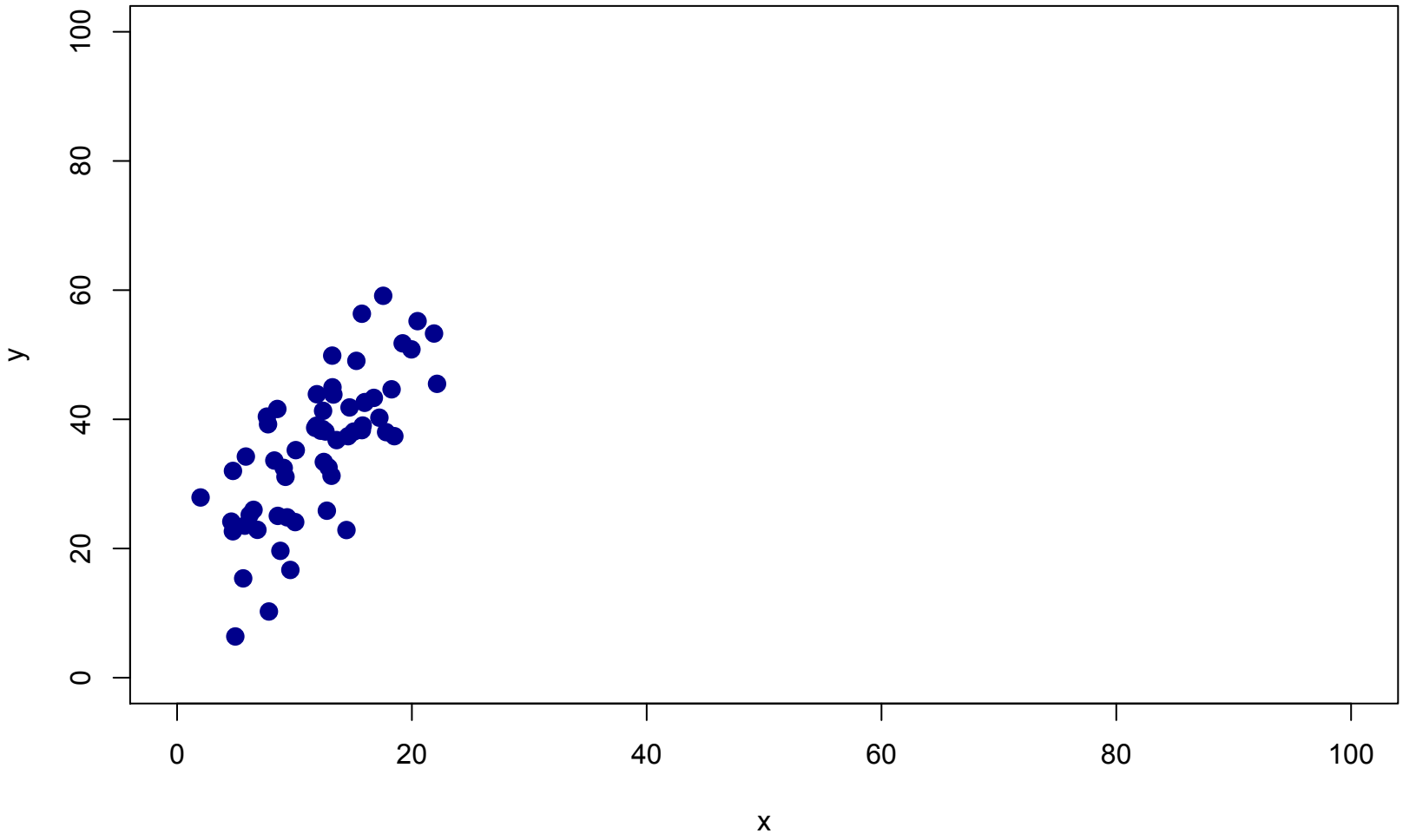
What is an outlier?

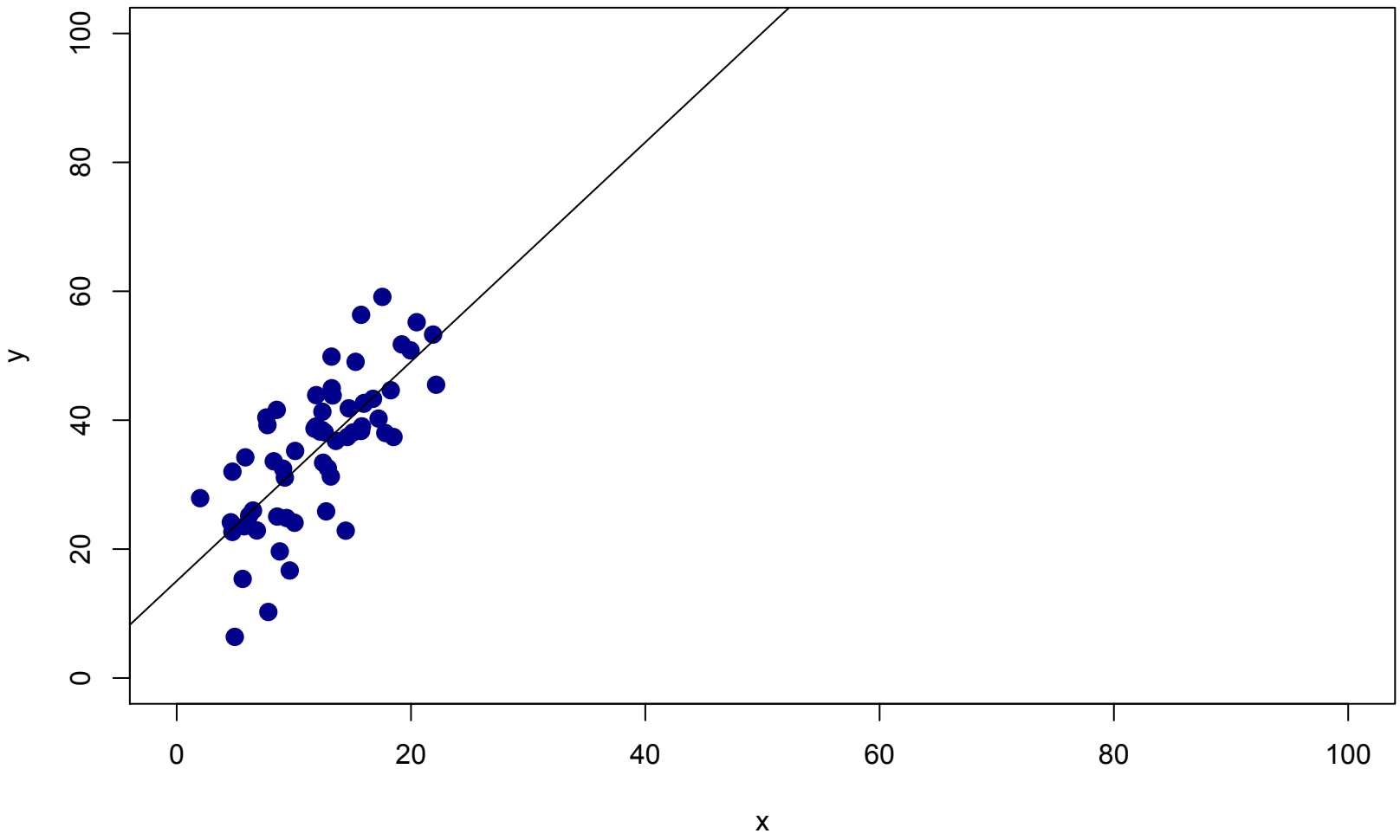
How big is a “big residual”?

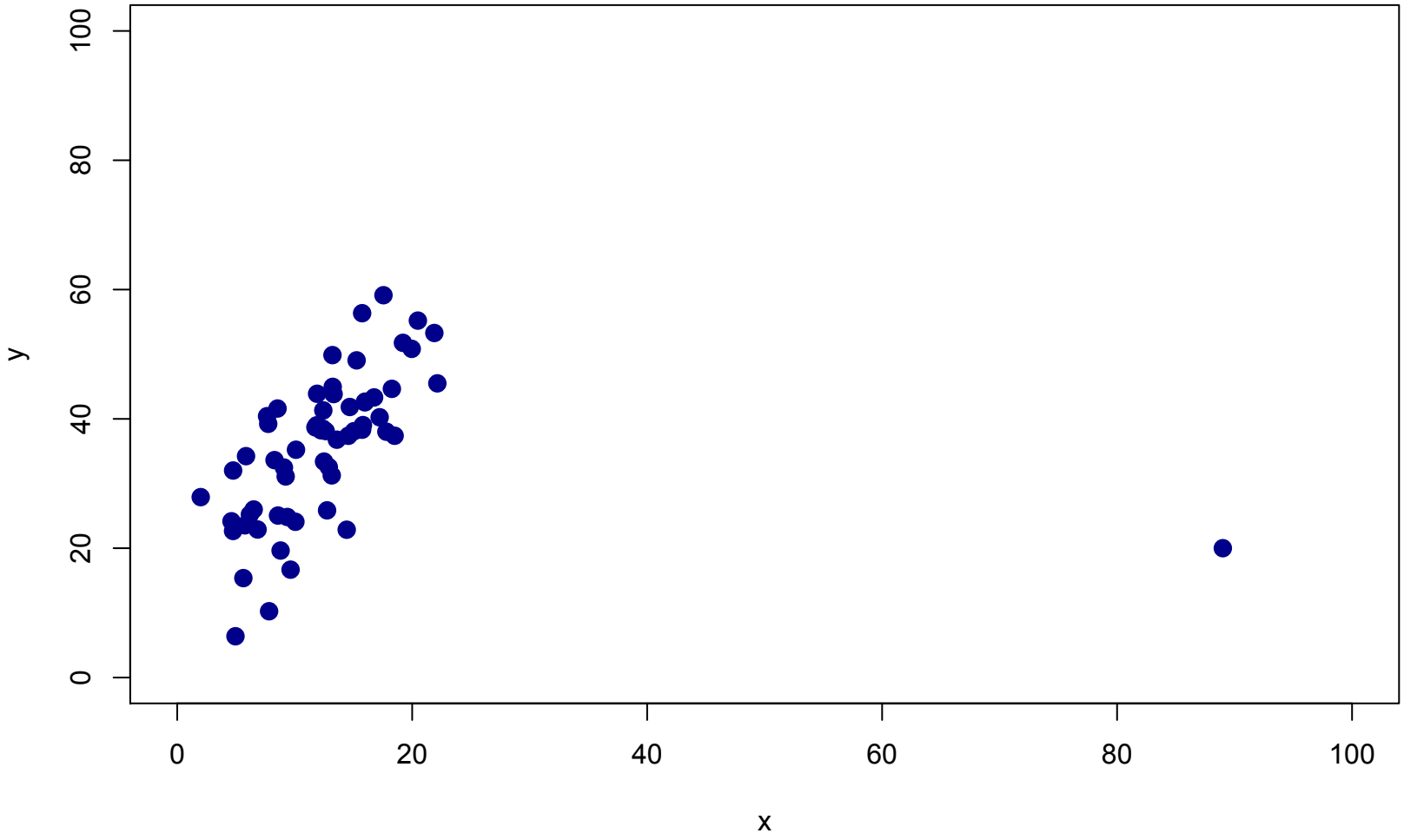
$$\epsilon_i \sim Normal(0, \sigma^2)$$

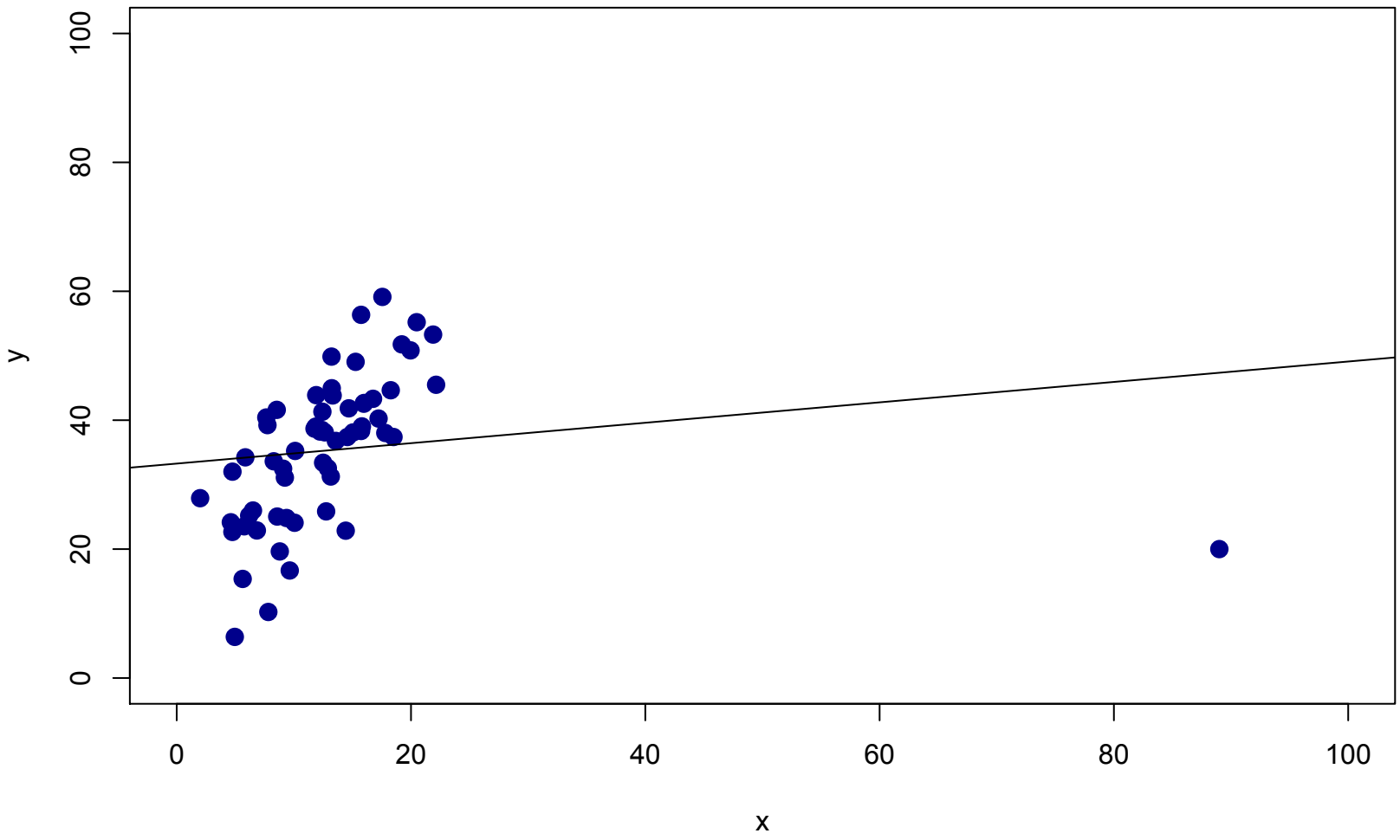
So maybe we could say that a big residual is…

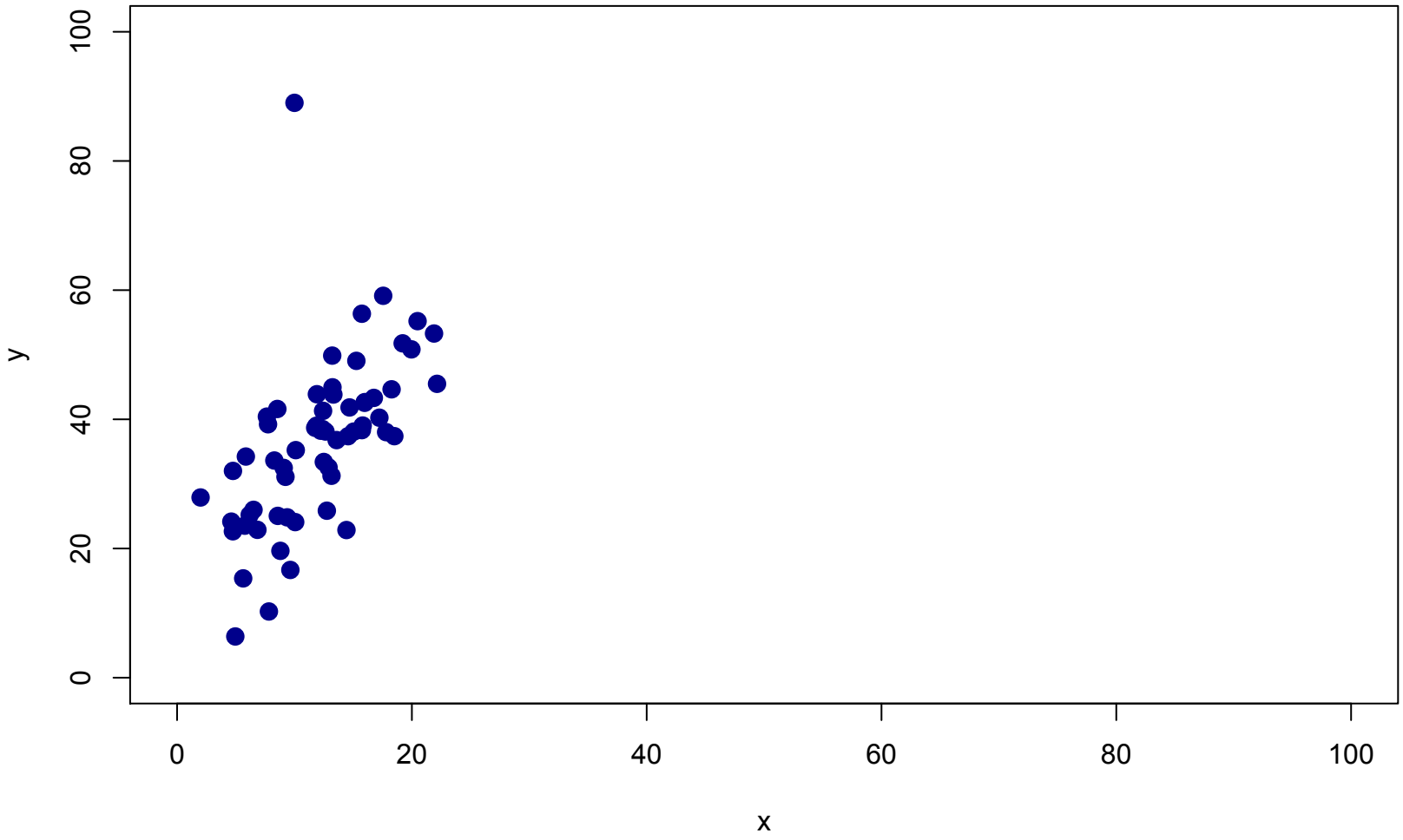
$\hat{\epsilon}_i > 2\hat{\sigma}$

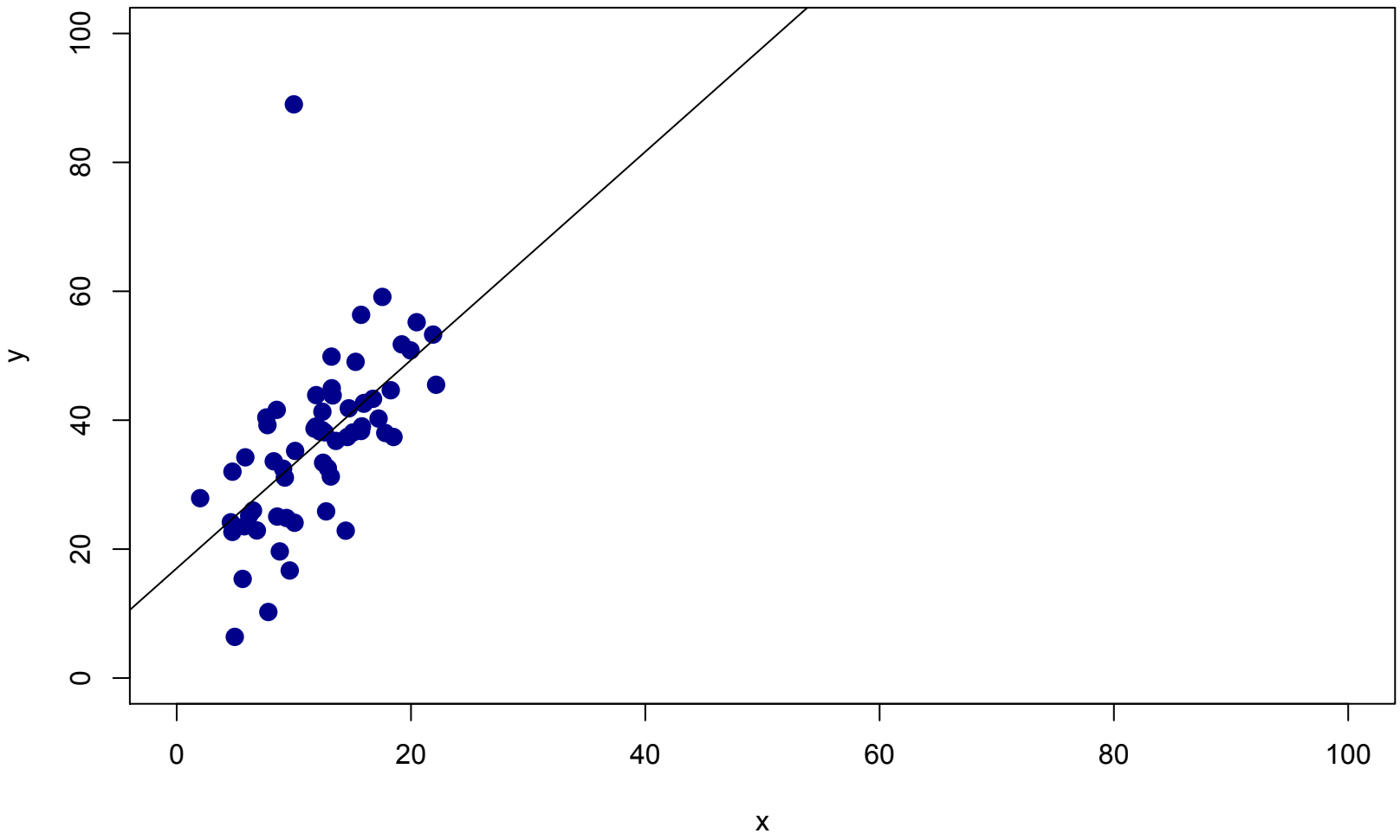

y
0
20
40
60
80
100
x
0
20
40
60
80
100

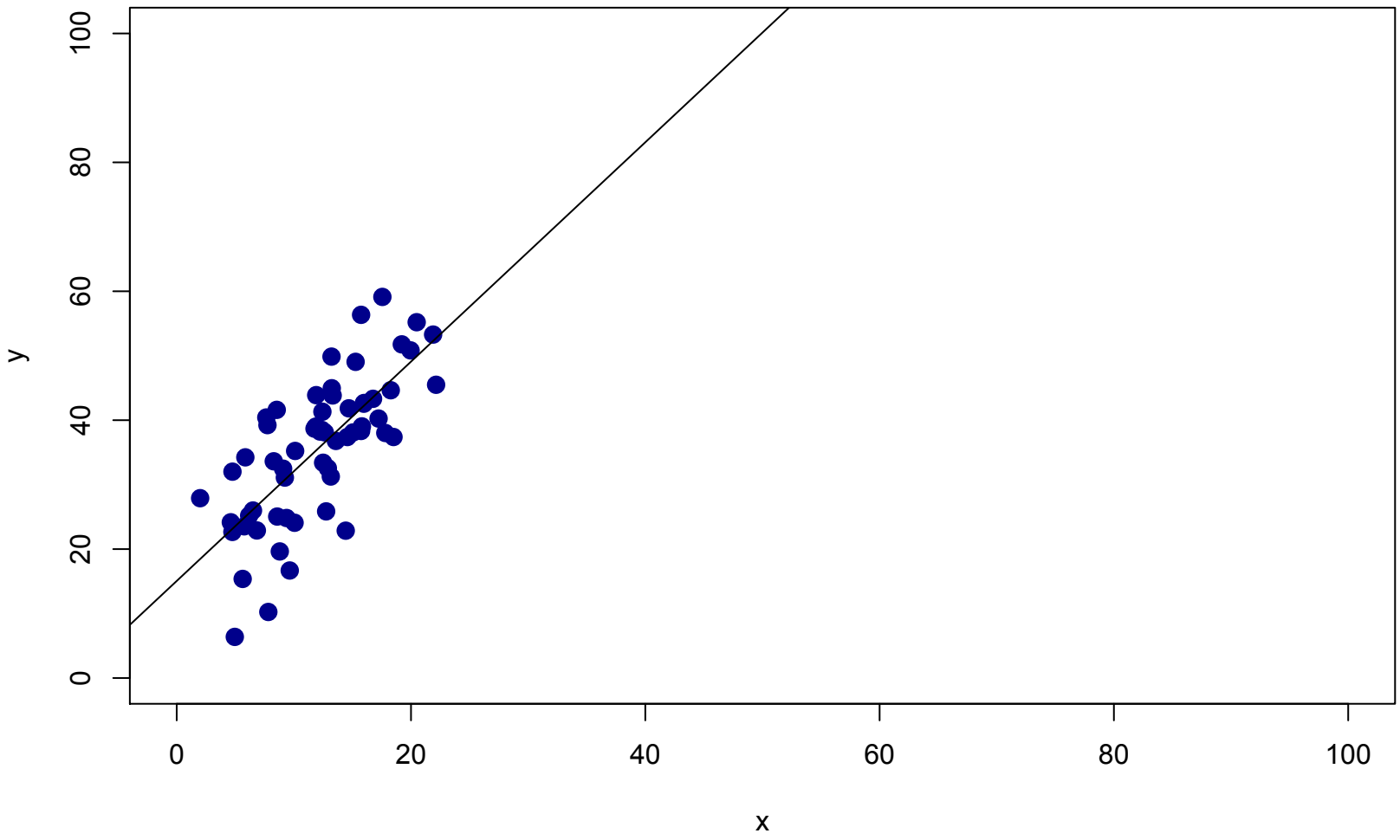

y
0
20
40
60
80
100
x
0
20
40
60
80
100

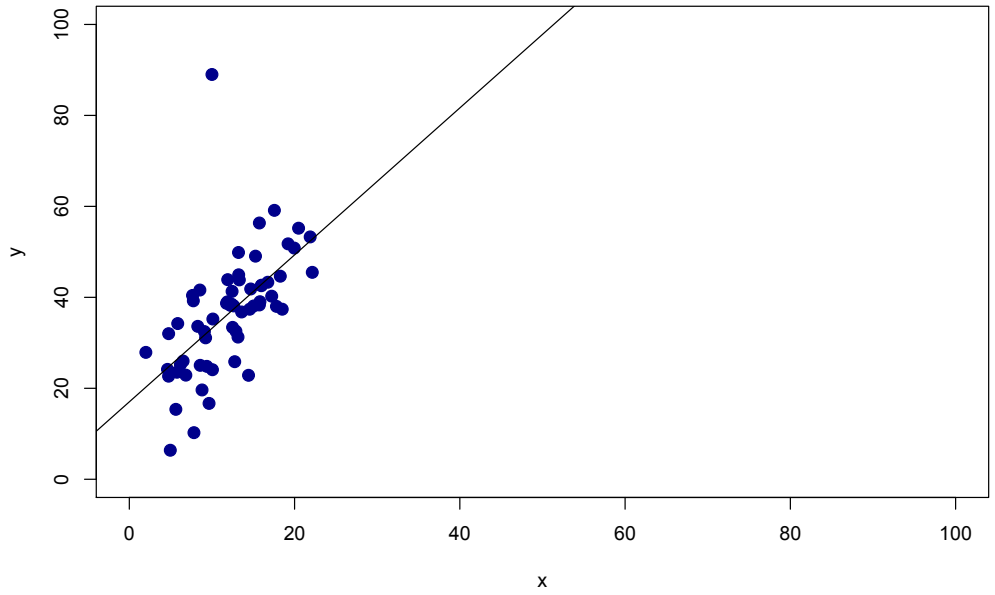

y
0
20
40
60
80
100
x
0
20
40
60
80
100

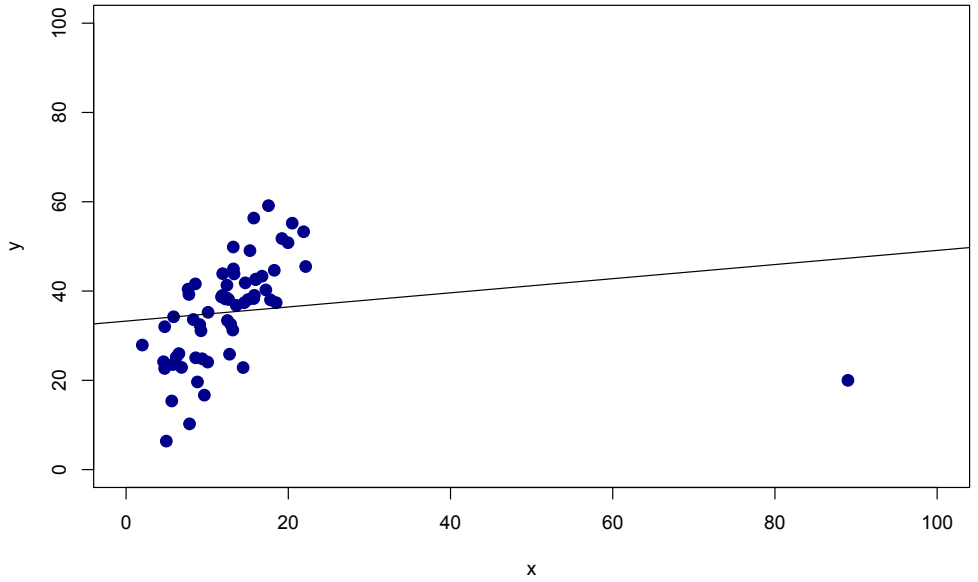

y
0
20
40
60
80
100
x
0
20
40
60
80
100


y
0
20
40
60
80
100
x
0
20
40
60
80
100


y
0
20
40
60
80
100
x
0
20
40
60
80
100


y
0
20
40
60
80
100
x
0
20
40
60
80
100


y
0
20
40
60
80
100
x
0
20
40
60
80
100


y
0
20
40
60
80
100
x
0
20
40
60
80
100

Hat matrix (aka Projection matrix)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y},$$

so the fitted values are

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}.$$

Therefore, the projection matrix (and hat matrix) is given by

$$\mathbf{P} \equiv \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}.$$

Hat matrix (aka Projection matrix)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y},$$

so the fitted values are

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}.$$

Therefore, the projection matrix (and hat matrix) is given by

$$\mathbf{P} \equiv \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}.$$

Therefore: $\hat{\mathbf{y}} = \mathbf{P}\mathbf{y}$

Hat matrix (aka Projection matrix)

$$\mathbf{P} \equiv \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}.$$

Therefore: $\hat{\mathbf{y}} = \mathbf{P}\mathbf{y}$

```
> x1 <- c(82, 45, 71, 22, 29, 9, 12, 18, 24)
> y <- c(71, 54, 43, 45, 21, 11, 30, 45, 10)
> n <- 9
> p<-1
> k<-2
>
> X<-cbind(1,x1)
>
> P<-X%*%solve(t(X)%*%X)%*%t(X)
> P%*%y
          [,1]
 [1,] 62.61099
 [2,] 42.33057
 [3,] 56.58167
 [4,] 29.72382
 [5,] 33.56066
 [6,] 22.59827
 [7,] 24.24263
 [8,] 27.53134
 [9,] 30.82006
>
> mod<-lm(y~x1)
> mod$fitted
       1        2        3        4        5        6
62.61099 42.33057 56.58167 29.72382 33.56066 22.59827
       7        8        9
24.24263 27.53134 30.82006
```

Hat diagonal

P_{ii}

The hat diagonal, P_{ii} , is a good measure of how much **influence** the ith observation has on the fitted model.

Hat diagonal

P_{ii}

The hat diagonal, P_{ii} , is a good measure of how much influence the ith observation has on the fitted model.

Big value of P_{ii} suggests that the ith observation is very *influential*.

How BiG?

Hat diagonal

P_{ii}

The hat diagonal, P_{ii}, is a good measure of how much influence the ith observation has on the fitted model.

Big value of P_{ii} suggests that the ith observation is very *influential*.

How BiG?

Note that:

$$\sum_{i=1}^{n} P_{ii} = k$$

Hat diagonal

P_{ii}

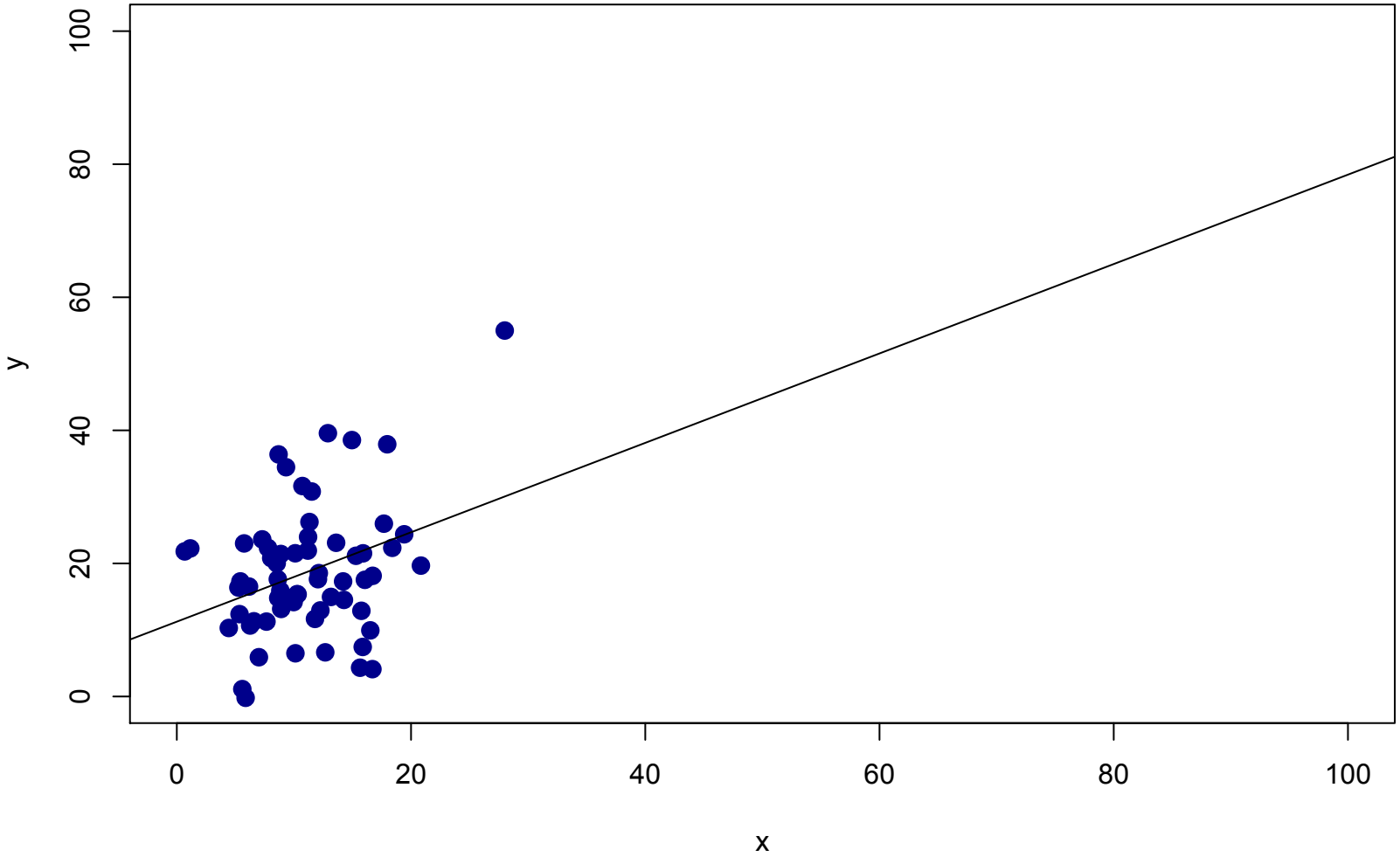
The hat diagonal, P_{ii}, is a good measure of how much influence the ith observation has on the fitted model.

Big value of P_{ii} suggests that the ith observation is very *influential*.

How BiG?

Note that: $$\sum_{i=1}^{n} P_{ii} = k$$

Therefore, the "average Pii" is equal to k/n. So think about "BiG" relative to average.


y
0
20
40
60
80
100
x
0
20
40
60
80
100

R code:

```
x<-(rnorm(60,0,5)+12)
y<-18+0.1*x+rnorm(60,0,8)

y[61]<-55
x[61]<-28

# Make the scatterplot:
plot(x,y, pch=20, col="darkblue", cex=2, axes=TRUE,ylim=c(0,100),
xlim=c(0,100), xlab="x", ylab="y")
abline(lm(y~x))

X<-cbind(1,x)
P<-X%*%solve(t(X)%*%X)%*%t(X)
diag(P)
round(diag(P),2)

k<-2
n<-61
k/n
mean(diag(P))
```

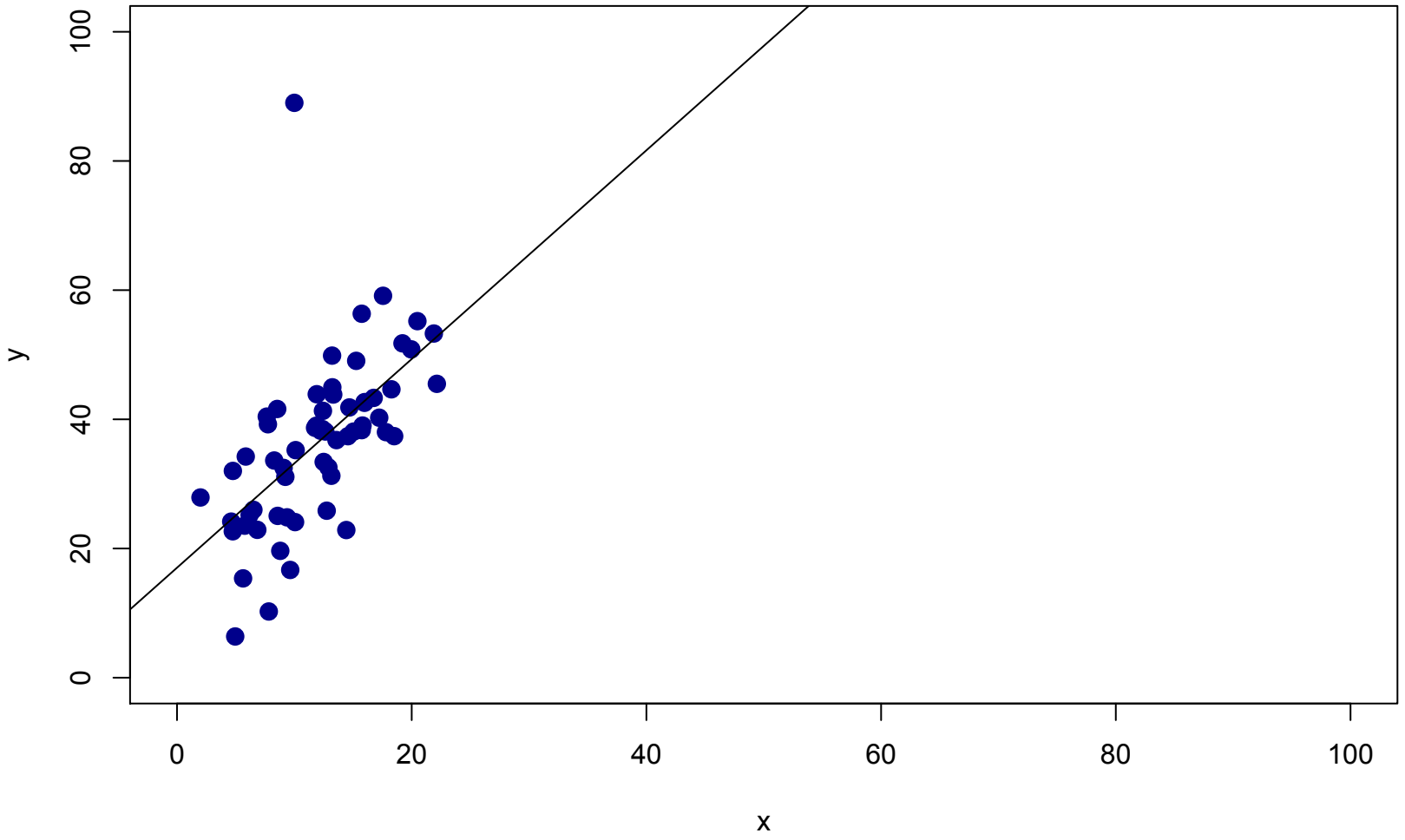
Remember three things:

Outliers are not necessarily influential (depends on leverage)

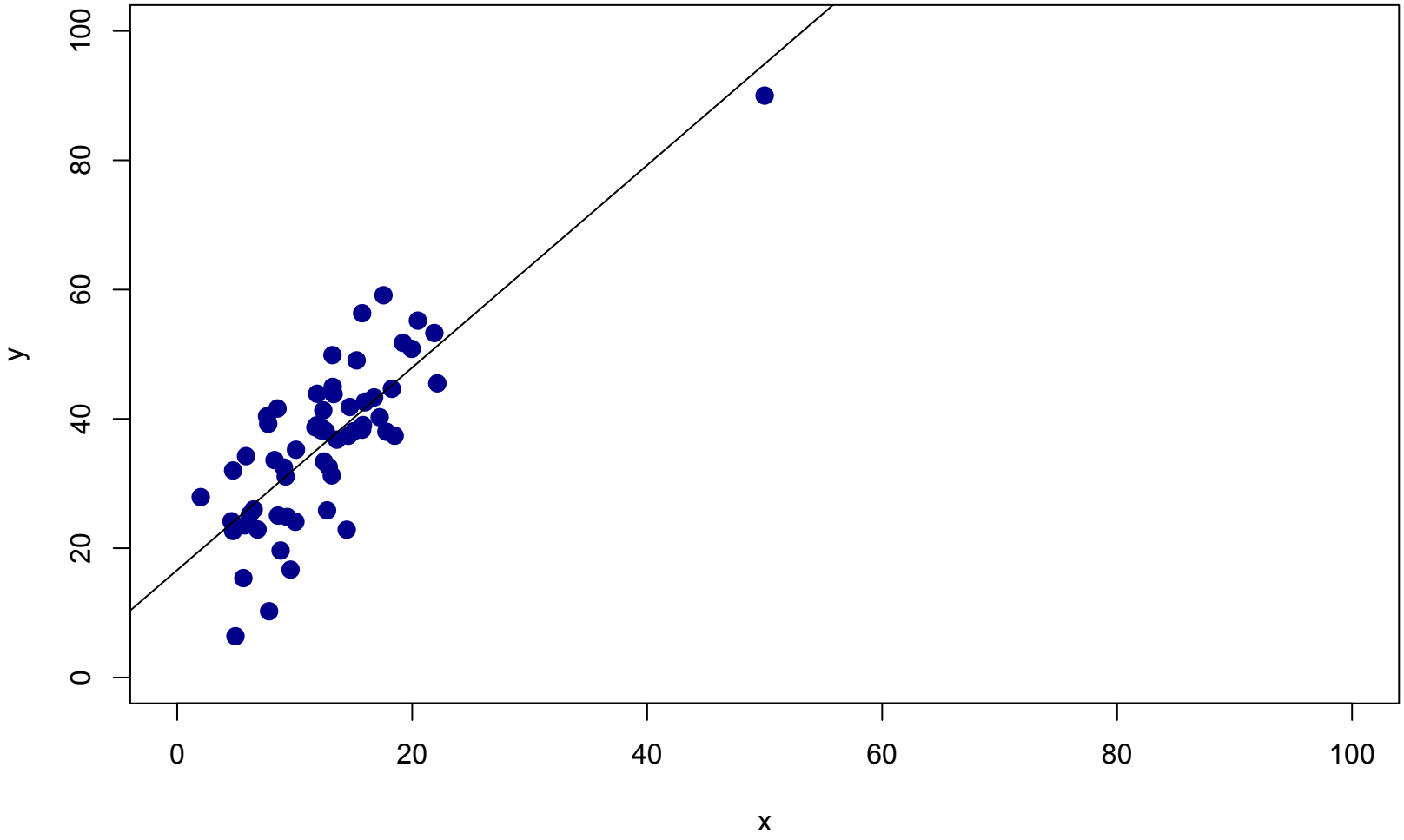
High leverage observations are not necessarily influential (is it an outlier?)

Influential points are not necessarily outliers

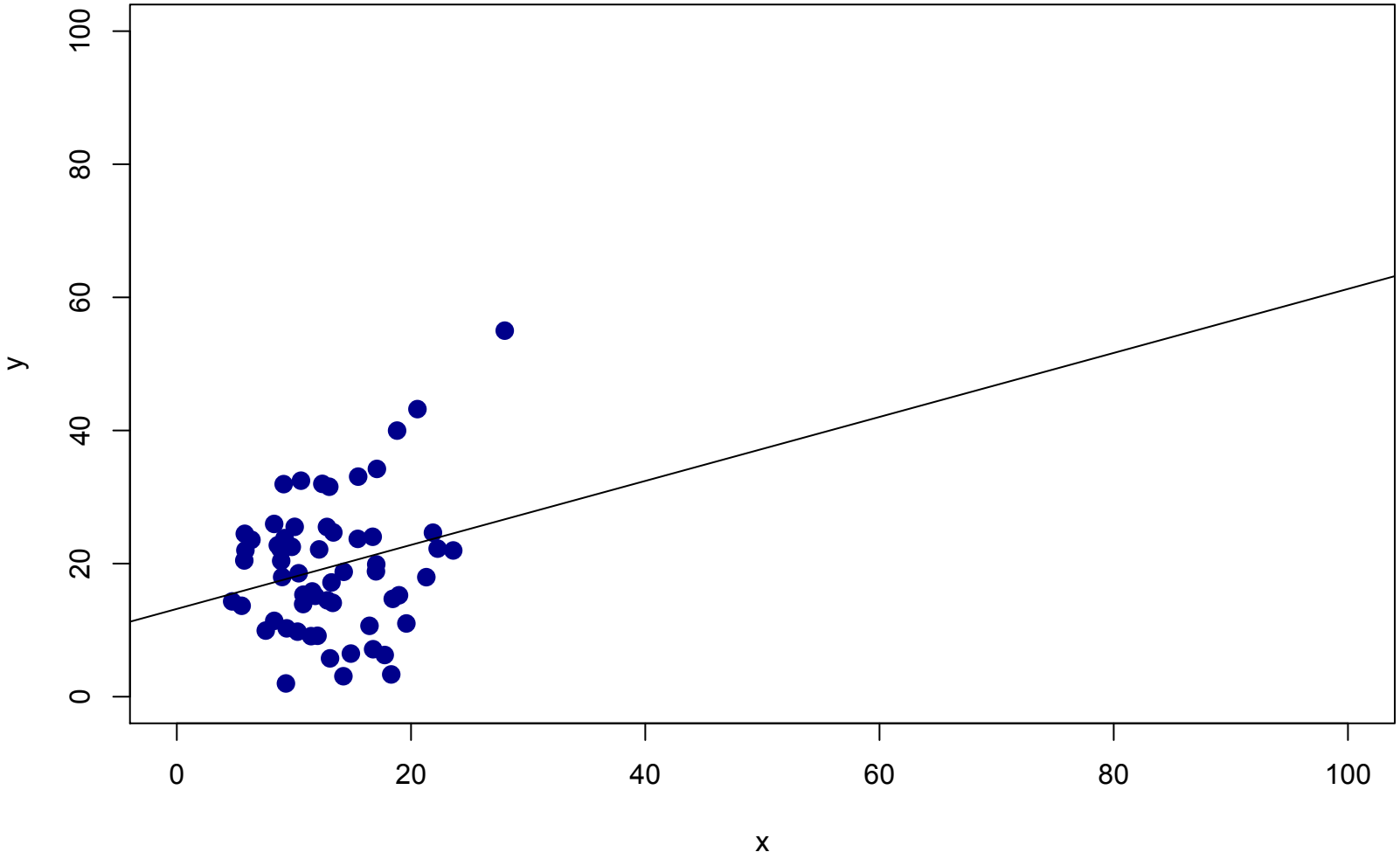
Outliers are not necessarily influential (depends on leverage)

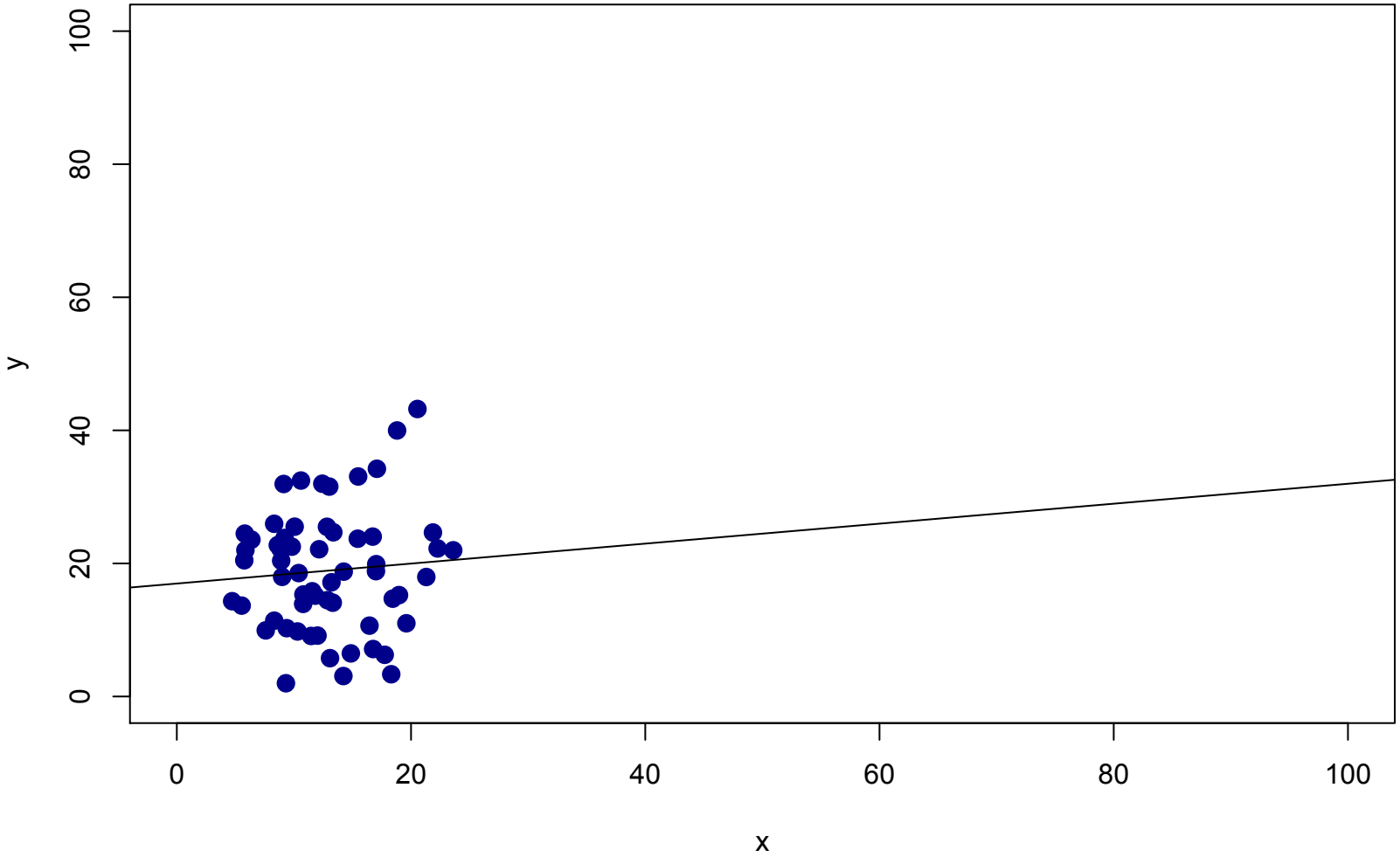


High leverage observations are not necessarily influential (is it an outlier?)



Influential points are not necessarily outliers




y
0
20
40
60
80
100
x
0
20
40
60
80
100

Also REMEMBER THAT:

HAT DIAGONALS INDICATE the *POTENTIAL* FOR BEING INFLUTENTIAL.

Variance of residuals is not equal!

The variances of the residuals at different X variable values may differ,
even if the variances of the errors at these different input variable values are equal.

Remember that there is a difference between errors and residuals:

The model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad , i = 1, ..., n$

The *fitted* model: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\epsilon}_i \quad , i = 1, ..., n$

errors: ϵ_i

residuals: $\hat{\epsilon}_i$

Variance of residuals is not equal!

Variance of residuals is not equal!

The variances of the residuals at different X variable values may differ,
even if the variances of the errors at these different input variable values are equal.

Remember that there is a difference between **errors** and **residuals**:

The model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad , i = 1, ..., n$

The *fitted* model: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\epsilon}_i \quad , i = 1, ..., n$

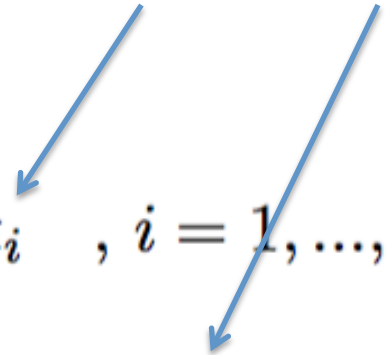
errors: $\epsilon_i \sim Normal(0, \sigma^2)$

residuals: $\hat{\epsilon}_i \sim \quad ???$

Variance of residuals is not equal!

The variances of the residuals at different X variable values may differ,
even if the variances of the errors at these different input variable values are equal.

Remember that there is a difference between **errors** and **residuals**:



The model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad , i = 1, ..., n$

The *fitted* model: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\epsilon}_i \quad , i = 1, ..., n$

errors: $\epsilon_i \sim Normal(0, \sigma^2)$

residuals: $\hat{\epsilon}_i \sim \quad ???$

Variance of residuals is not equal!

The variances of the residuals at different X variable values may differ,
even if the variances of the errors at these different input variable values are equal.

Remember that there is a difference between errors and residuals:

The model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad , i = 1, ..., n$

The *fitted* model: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\epsilon}_i \quad , i = 1, ..., n$

errors: $\epsilon_i \sim Normal(0, \sigma^2)$

residuals: $\hat{\epsilon}_i \sim Normal(0, \sigma^2(1 - P_{ii}))$

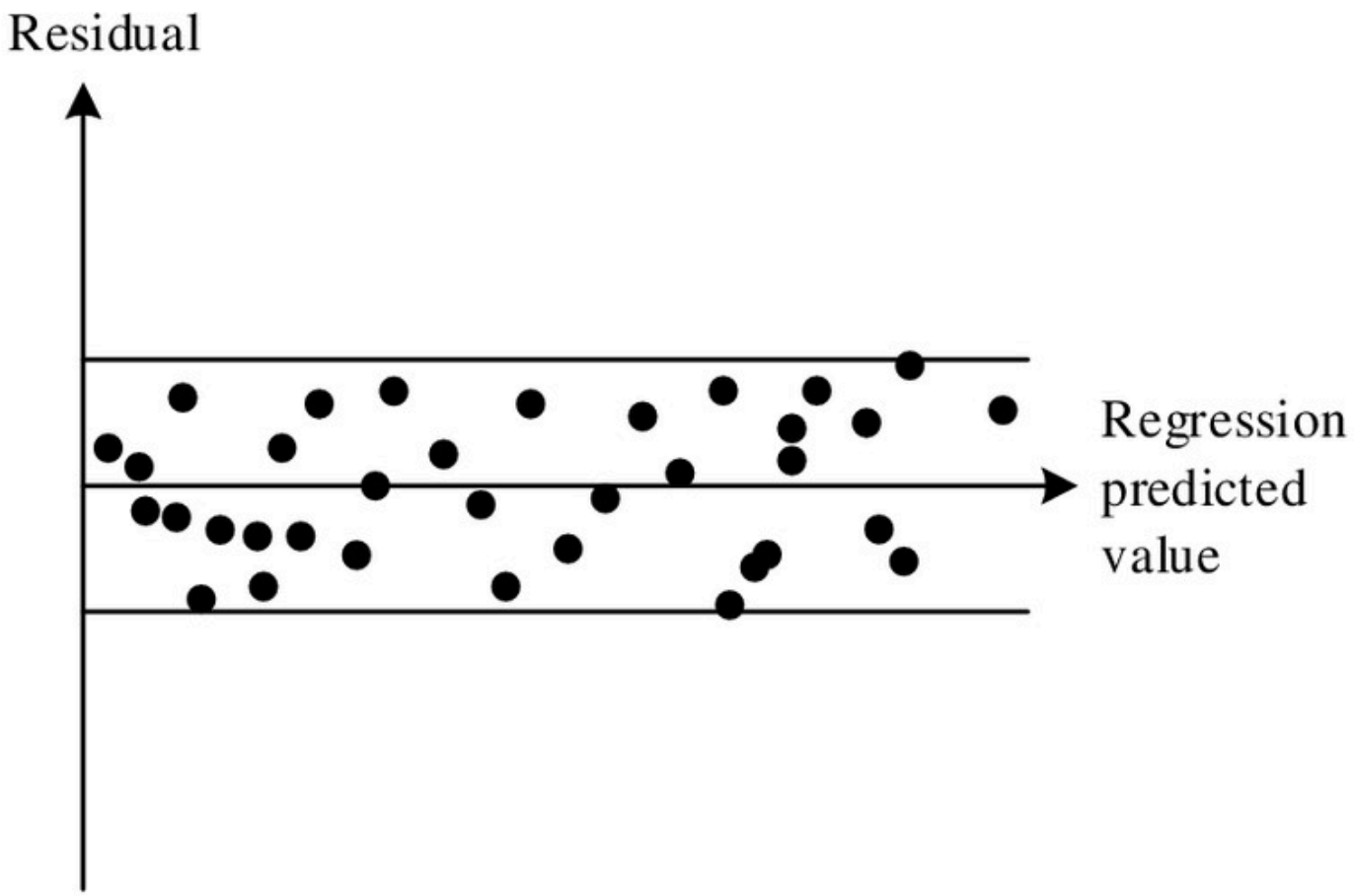
Variance of residuals is not equal!

Variance-Covariance Matrix of the residuals:

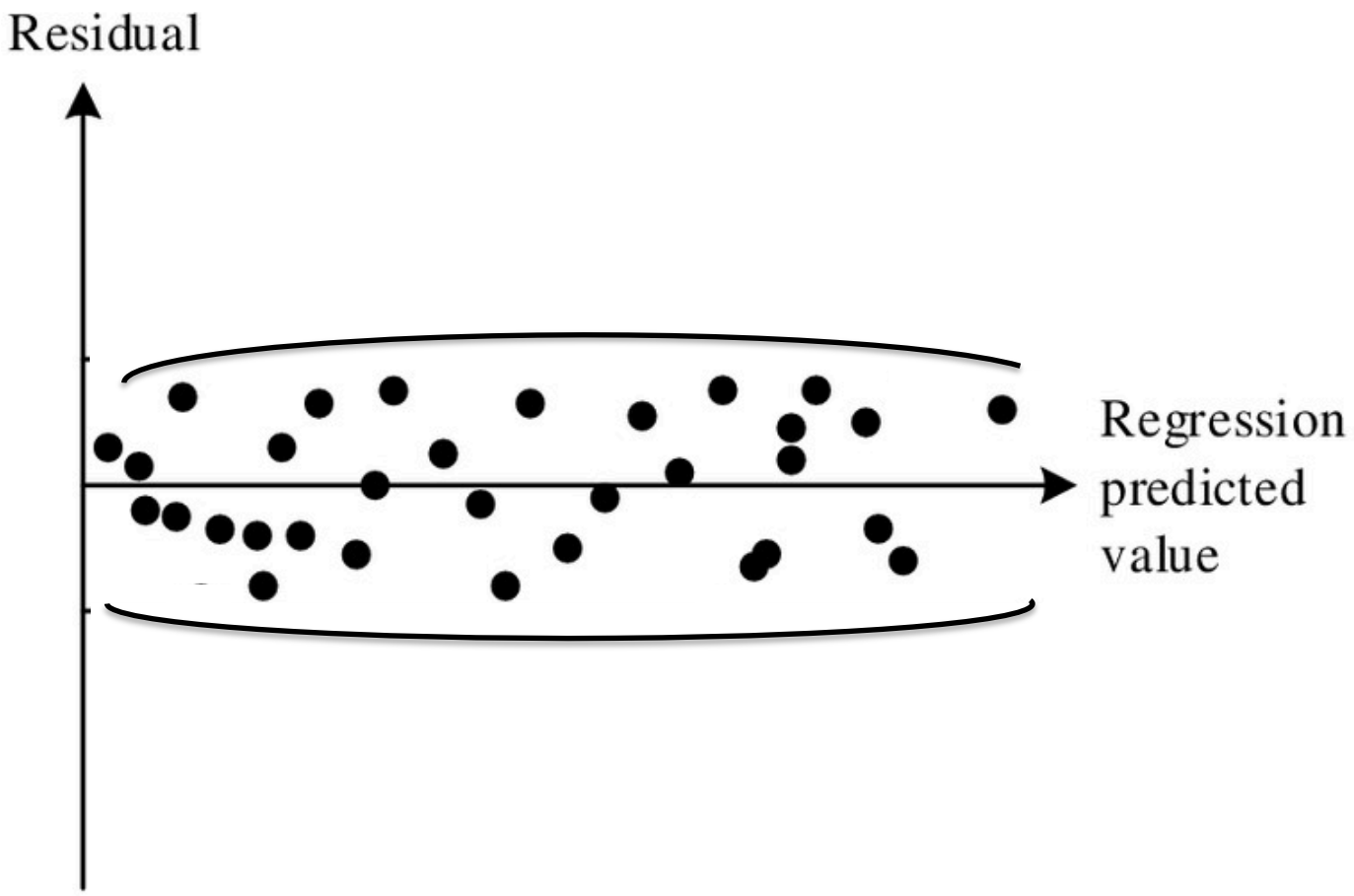
$$Var(\hat{\epsilon}) = \sigma^2 \begin{pmatrix} 1-h_{11} & -h_{12} & -h_{13} & \cdots & \cdots & -h_{1n} \\ -h_{21} & 1-h_{22} & -h_{23} & \cdots & \cdots & -h_{2n} \\ \cdots & \cdots & \ddots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \ddots & \cdots & \cdots \\ -h_{n1} & -h_{n2} & -h_{n3} & \cdots & \cdots & 1-h_{nn} \end{pmatrix}$$

What does this mean for our residual diagnostic plots?

Residual plot (homoscedasticity)



Residual plot (homoscedasticity)



```
variance_of_residuals<-(summary(lm(y~x))$sigma^2)*(1-diag(P))

plot(variance_of_residuals~x)

plot(lm(y~x)$residuals)

plot(lm(y~x)$residuals/sqrt(variance_of_residuals))

lm(y~x)$residuals/sqrt(variance_of_residuals)

ls.diag(lm(y~x))$std.res
```

What does this mean for our residual diagnostic plots?

Maybe we should adjust the residuals before plotting them…

Output of ls.diag() in R [diagnostics after lsfit or lm].

See Section 4.3 for full details.

An observation (row of data matrix) is influential if the value of $\hat{\boldsymbol{\beta}}$ changes a lot when this observation is deleted. First pass through these notes: influential observations only, as ideas are related to cross-validation with leave-one-out.

```
[1] "std.dev" "hat"         "std.res"  "stud.res"   "cooks"
[6] "dfits"   "correlation"  "std.err"  "cov.scaled" "cov.unscaled"
```

n =sample size, $\mathbf{X}$=data matrix of explanatory variables of dimension $n \times k$ with first column of 1s for the intercept.

---

1. `std.dev`: residualSD $= \hat{\sigma} = \sqrt{\sum_i e_i^2/(n-k)}$

1. `std.dev`: residualSD $= \hat{\sigma} = \sqrt{\sum_i e_i^2/(n-k)}$

2. `hat`: diagonal of projection or hat matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, ith diagonal element denoted as P_{ii}.

1. `std.dev`: residualSD $= \hat{\sigma} = \sqrt{\sum_i e_i^2/(n-k)}$

2. `hat`: diagonal of projection or hat matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, ith diagonal element denoted as P_{ii}.

3. `std.res`: vector of standardized residuals: $e_i^* = e_i/[\hat{\sigma}\sqrt{1-P_{ii}}]$

standardized residuals:
these are residuals normalized to unit variance.

these are residuals divided by their standard error.

1. `std.dev`: residualSD $= \hat{\sigma} = \sqrt{\sum_i e_i^2/(n-k)}$

2. `hat`: diagonal of projection or hat matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, ith diagonal element denoted as P_{ii}.

3. `std.res`: vector of standardized residuals: $e_i^* = e_i/[\hat{\sigma}\sqrt{1-P_{ii}}]$

standardized residuals:
these are residuals normalized to unit variance.

these are residuals divided by their standard error.

```
> mod$residuals/(summary(mod)$sigma*sqrt(1-diag(P)))
```

1. `std.dev`: residualSD $= \hat{\sigma} = \sqrt{\sum_i e_i^2/(n-k)}$

2. `hat`: diagonal of projection or hat matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, ith diagonal element denoted as P_{ii}.

3. `std.res`: vector of standardized residuals: $e_i^* = e_i/[\hat{\sigma}\sqrt{1-P_{ii}}]$

1. `std.dev`: residualSD $= \hat{\sigma} = \sqrt{\sum_i e_i^2/(n-k)}$

2. `hat`: diagonal of projection or hat matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, ith diagonal element denoted as P_{ii}.

3. `std.res`: vector of standardized residuals: $e_i^* = e_i/[\hat{\sigma}\sqrt{1-P_{ii}}]$

4. `stud.res`: vector of Studentized residuals

studentized residuals:
these are residuals normalized to unit variance…

where the estimate of the variance is done without the i^{th} observation.

4. stud.res: vector of Studentized residuals

studentized residuals:
these are residuals normalized to unit variance…

where the estimate of the variance is done
without the i^{th} observation.

```
> s_without_i <- rep(0,length(y))
>
> for(i in 1:length(y)){
+   s_without_i[i] <- summary(lm(y[-i]~x1[-i]))$sigma
+ }
>
> studentized_residuals<-mod$residuals/(s_without_i*sqrt(1-diag(P)))
>
```

4. `stud.res`: vector of Studentized residuals

studentized residuals:
these are residuals normalized to unit variance…

where the estimate of the variance is done without the i^{th} observation.

The **"estimate of the variance is done without the i^{th} observation"** can be done more easily:

$$s_{-i} = \sqrt{\frac{(n-k)s^2 - \hat{\epsilon}_i^2/(1-P_{ii})}{n-k-1}}$$

So many kinds of residuals!!!

The difference between **standardized** and **studentized** residuals is often very small or even negligible.

The difference between the two will depend on the **amount of influence** of the observation has on the model fit.

5. `cooks`: Cook's distance = vector of inverse-covariance-matrix weighted squared distances of $\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}}$ to measure influence of the observations:

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}})^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}})}{k\hat{\sigma}^2} = \frac{(e_i^*)^2}{k} \times \frac{P_{ii}}{1 - P_{ii}},$$

$\hat{\boldsymbol{\beta}}_{-i}$ is the vector of regression coefficients with $(\mathbf{x}_i, y_i)$ omitted. D_i is a distance measure that is invariant to scaling of explanatory variables; also other invariances.

where:

$$e_i^* = e_i / [\hat{\sigma}\sqrt{1 - P_{ii}}]$$

BASIC IDEA:

"Cook's distance is the sum of all the changes in the fitted values of a regression model when the i^{th} observation is removed from the data"

6. `dfits`: Another measure of influence of the ith observation: $\text{dfits}_i = (\hat{y}_i - \hat{y}_{i|-i})/[\hat{\sigma}_{-i}\sqrt{P_{ii}}]$.

7. `correlation`: $V = (\mathbf{X}^T\mathbf{X})^{-1}$ converted to a correlation matrix, that is, $v_{ij} \to v_{ij}/\sqrt{v_{ii}v_{jj}}$.

6. `dfits`: Another measure of influence of the ith observation: $\text{dfits}_i = (\hat{y}_i - \hat{y}_{i|-i})/[\hat{\sigma}_{-i}\sqrt{P_{ii}}]$.

7. `correlation`: $V = (\mathbf{X}^T\mathbf{X})^{-1}$ converted to a correlation matrix, that is, $v_{ij} \rightarrow v_{ij}/\sqrt{v_{ii}v_{jj}}$.

8. `std.err`: vector of SEs of the $\hat{\beta}_j$ or the square roots of the diagonal of $\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$.

9. `cov.scaled`: $\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$ = estimated covariance matrix of $\hat{\boldsymbol{\beta}}$.

10. `cov.unscaled`: $(\mathbf{X}^T\mathbf{X})^{-1}$

What is an outlier?

How big is a “big residual”?

$$\text{residuals: } \hat{\epsilon}_i \sim Normal(0, \sigma^2(1 - P_{ii}))$$

So to answer our question we should calculate Studentized or Standardized residuals.

A Rule of Thumb is that a Studentized or Standardized residual:

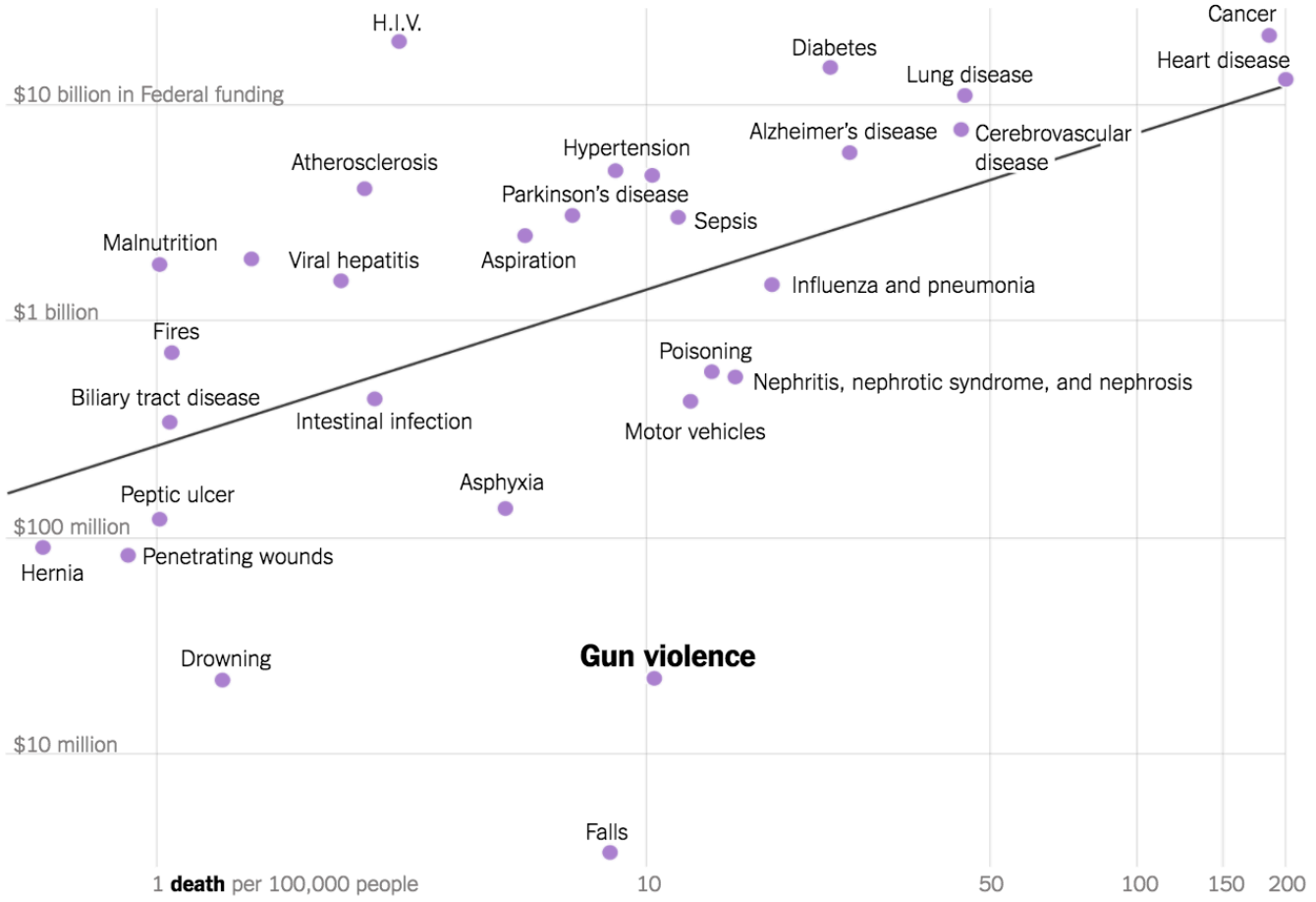
- larger than 2 is **BIG** (“outlying”) and
- larger than 3 is **VERY BIG** (“outlier”, you should check to make sure there is not a mistake)

The New York Times

There's an Awful Lot We Still Don't Know About Guns

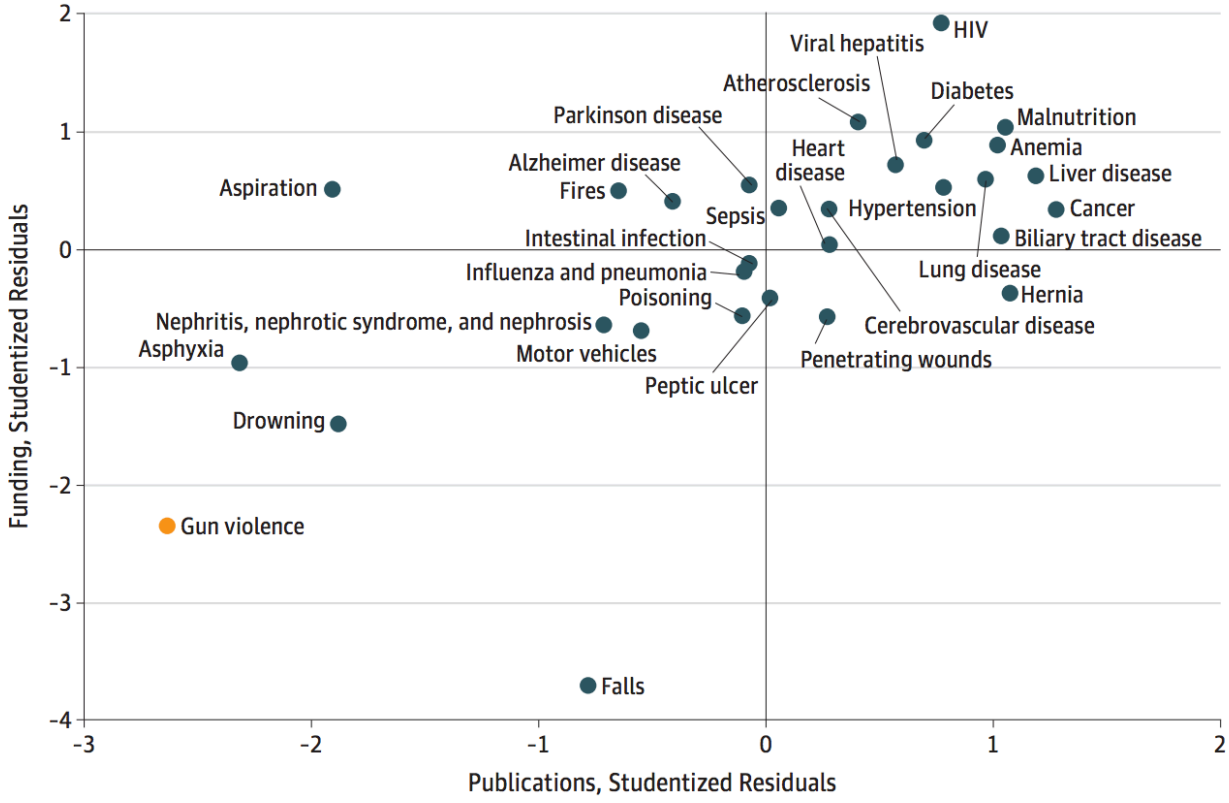
By **QUOCTRUNG BUI** and **MARGOT SANGER-KATZ** MARCH 2, 2018

Federal funding for research on leading causes of death



Source: From 2004 to 2014, David Stark and Nigam Shah, Funding and Publication of Research on Gun Violence and Other Leading Causes of Death

Figure 2. Studentized Residual Predicted vs Observed Funding and Publication Volumes for 30 Leading Causes of Death in the United States



HIV indicates human immunodeficiency virus. Mortality rate was used to predict funding and research volume. Studentized residuals (residual divided by estimated standard error) were calculated to give a standardized estimate of predicted vs observed funding and publication volume. The 4 quadrants represent observed funding greater than predicted, observed publication volume less than predicted (upper-left); observed funding and publication volume greater than predicted (upper-right); observed funding less than predicted, observed publication volume greater than predicted (lower-right); observed funding and publication volume less than predicted (lower-left).

BONUS FOR LEAVE-ONE-OUT Cross Validation

The cross-validated root mean square (prediction) error is:

$$CVRMSE_{\text{leaveoneout}}(x_1, \ldots, x_p) = \sqrt{n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_{i|-i})^2} \,. \tag{4.26}$$

$CVRMSE_{\text{leaveoneout}}(\mathbf{x}_J)$ can be also defined for the subset of the explanatory variables indexed by $J \subset \{1, \ldots, p\}$.

BONUS FOR LEAVE-ONE-OUT Cross Validation

The cross-validated root mean square (prediction) error is:

$$CVRMSE_{\text{leaveoneout}}(x_1, \ldots, x_p) = \sqrt{n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_{i|-i})^2}. \tag{4.26}$$

$CVRMSE_{\text{leaveoneout}}(\mathbf{x}_J)$ can be also defined for the subset of the explanatory variables indexed by $J \subset \{1, \ldots, p\}$.

It turns out there is a simple formula for cross-validation leave-one-out residuals (without having to compute n different regressions). An identity is

$$y_i - \hat{y}_{i|-i} = (y_i - \hat{y}_i)/(1 - P_{ii}), \quad P_{ii} = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i, \tag{4.27}$$

where $\mathbf{x}_i^T = (1, x_{i1}, \ldots, x_{ip})$. Alternatively, P_{ii} is the ith diagonal element of the projection matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$. See Section 4.3 for why this quantity is called the projection matrix.