Limits of linear regression

Readings for today

- Chapter 3: Linear regression. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: with applications in R (Vol. 6). New York: Springer.

Topics

1. Interpretation constraints

2. Pitfalls of linear regression

Interpretation constraints

Goals of learning f

Intuition

1. Inference: understand how changes in X associate with changes in Y.

2. Prediction: predict a future observation in Y from X.

Control

(James et al. 2013)

Example: Inference

Q: What factors associate with house value?

Model:

$$Y_\$ = \hat{\beta}_0 + \hat{\beta}_{crime} X_{crime} + \hat{\beta}_{age} X_{age} + \hat{\beta}_{tax} X_{tax} + \hat{\beta}_{s:t} X_{s:t}$$

	$\hat{\beta}_i$	$SE(\hat{\beta}_i)$	*t*	*p*	*sig*
Intercept	36.46	5.10	7.14	3.28E-12	***
Crime	-1.08	0.33	-3.29	0.002	**
Age	0.01	0.01	0.052	0.85	
Tax	-0.12	0.01	-3.28	0.001	**
S:T ratio	-0.95	0.13	-7.28	0.0005	***

Variables:

- $Y_\$$: median house value
- X_{crime}: local crime rate
- X_{age}: # houses over 80 years old
- X_{tax}: local property tax rate
- $X_{s:t}$: school student-teacher ratio

(James et al. 2013)

Example: Prediction

Q: What factors predict house value?

Model:

$$Y_\$ = \hat{\beta}_0 + \hat{\beta}_{crime}X_{crime} + \hat{\beta}_{age}X_{age} + \hat{\beta}_{tax}X_{tax} + \hat{\beta}_{s:t}X_{s:t}$$

Steps:

1. Fit $\hat{f}_{train}(X_{train})$
2. Predict $\hat{Y}_{test}$ using $\hat{f}_{train}(X_{test})$
3. Evaluate test error:

$$\sum_{i=1}^{n} (y_i^{test} - \hat{y}_i^{test})^2$$

$$\begin{array}{c} Y_{train}, X_{train} \\ \\ Y_{test}, X_{test} \end{array}
\begin{pmatrix} y_i \\ \vdots \\ y_m \\ \hline y_{m+1} \\ \vdots \\ y_n \end{pmatrix}
=
\begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} & x_{1,5} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{m,1} & x_{m,2} & x_{m,3} & x_{m,4} & x_{m,5} \\ \hline x_{m+1,1} & x_{m+1,2} & x_{m+1,3} & x_{m+1,4} & x_{m+1,5} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & x_{n,3} & x_{n,4} & x_{n,5} \end{pmatrix}
\begin{array}{l} \text{Training set} \\ \\ \text{Test set} \end{array}$$

(James et al. 2013)

Qualitative predictors

Q: Are there gender differences in reading comprehension?

Variables:

- Y_{read}: content recall score
- X_{gender}: gender (male, female)

Model:

$$Y_{read} = \hat{\beta}_0 + \hat{\beta}_1 X_{gender}$$

Case 1:

$$X_{gender} = \begin{cases} 0, & \text{male} \\ 1, & \text{female} \end{cases}$$

$$\text{male} = \hat{\beta}_0$$

$$\text{female} = \hat{\beta}_1$$

Case 2:

$$X_{gender} = \begin{cases} -1, & \text{male} \\ 1, & \text{female} \end{cases}$$

$$\text{male} = \hat{\beta}_0 - \hat{\beta}_1$$

$$\text{female} = \hat{\beta}_0 + \hat{\beta}_1$$

How you specify your qualitative predictor variables determines how you interpret the effects

(James et al. 2013)

Qualitative predictors (>2 levels)

Q: Are there gender differences in reading comprehension?

Variables:

- Y_{read}: content recall score
- X_{gender}: gender (male, female, nonbinary)

Dummy (binary) coding:

$$X_{female} = \begin{cases} 1, & \text{female} \\ 0, & \text{otherwise} \end{cases} \qquad X_{nb} = \begin{cases} 1, & \text{non-binary} \\ 0, & \text{otherwise} \end{cases}$$

$$\text{male} = \hat{\beta}_0$$

$$\text{female} = \hat{\beta}_0 + \hat{\beta}_2$$

$$\text{non-binary} = \hat{\beta}_0 + \hat{\beta}_1$$

Model:

$$Y_{read} = \hat{\beta}_0 + \hat{\beta}_1 X_{female} + \hat{\beta}_2 X_{nb}$$

(James et al. 2013)

Interactions

Q: Does gender influence the effect of age on reading comprehension?

Variables:

- Y_{read}: content recall score
- X_{gender}: gender (male, female)
- X_{age}: age (in years)

Model:

$$Y_{read} = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 X_{gender}}_{\text{main effect of gender}} + \underbrace{\hat{\beta}_2 X_{age}}_{\text{main effect of age}} + \underbrace{\hat{\beta}_3 X_{gender} X_{age}}_{\text{gender by age interaction}}$$

Hierarchical Principle: When including interaction terms, *always* include the main effect terms in the model.

(James et al. 2013)

Why need the Hierarchical Principle?

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2$$

$$= \hat{\beta}_0 + \underbrace{(\hat{\beta}_1 + \hat{\beta}_3 X_2)}_{\hat{\beta}_1^*} X_1 + \hat{\beta}_2 X_2$$

$$= \hat{\beta}_0 + \hat{\beta}_1^* X_1 + \hat{\beta}_2 X_2$$

- $\hat{\beta}_1^*$ depends on both X_1 and X_2 in order to describe the total relationship that X_1 has with Y.
- Excluding the main effect term $\hat{\beta}_1 X_1$ removes that portion of the variance explained by $\hat{\beta}_1^*$.

(James et al. 2013)

Pitfalls of linear regression

Structure of a linear model

Fundamental form:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p Xp + \epsilon$$

Solution:

$$\hat{\beta}_0 = E[Y] + \hat{\beta}_1 E[X]$$

$$= \bar{y} + \sum_{j=1}^{p} \hat{\beta}_j \bar{x}_p$$

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} (x_{i,j} - \bar{x}_j)(y_i - \bar{y})}{\sum_{i=1}^{n} (x_{i,j} - \bar{x}_j)^2}$$

Assumptions:

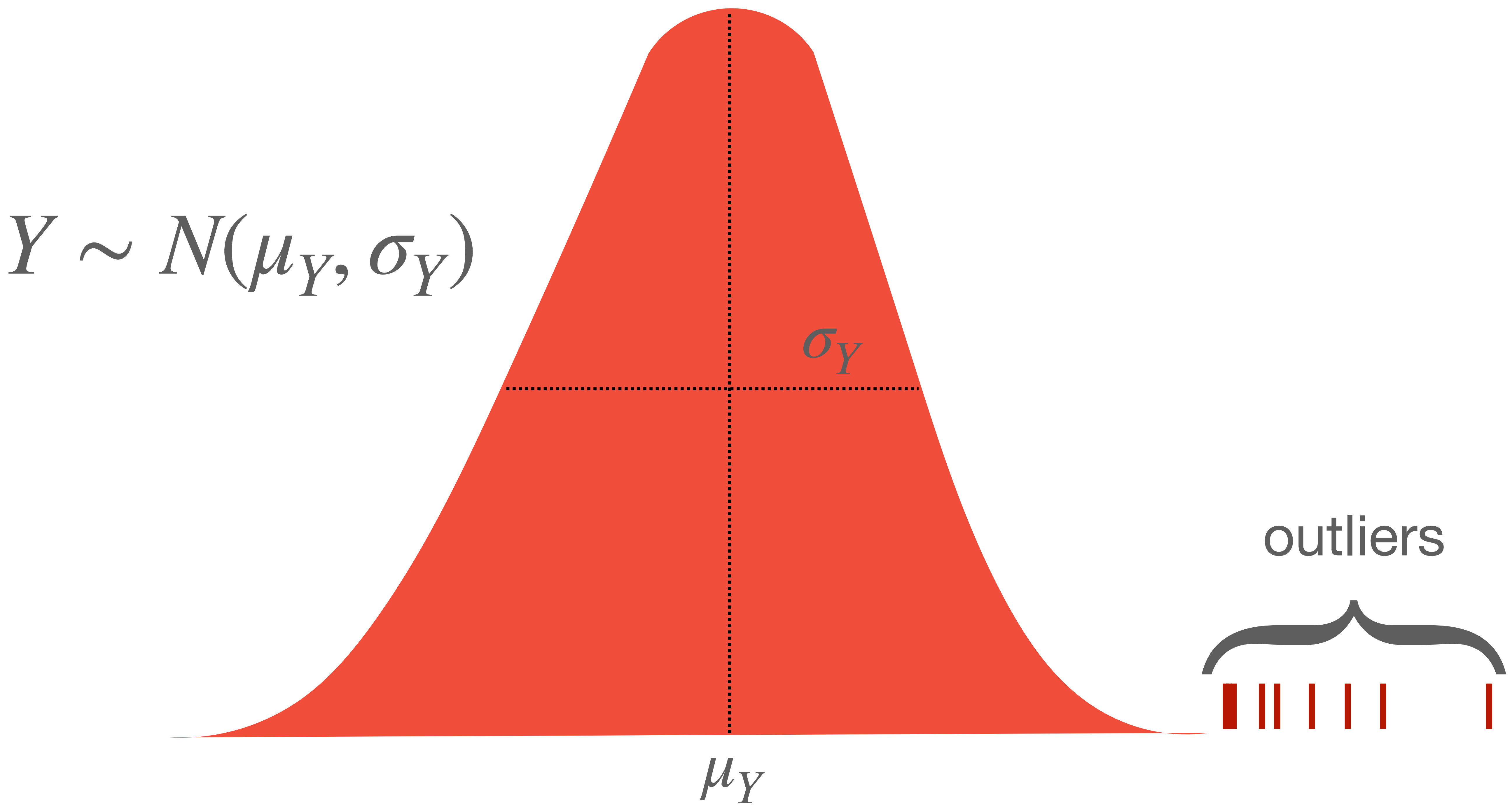
1. $f(X)$ describes a linear relationship between X and Y.
2. Y is normally distributed.
3. There is no collinearity between features in X.
4. $f(X)$ is stationary.

i.i.d.

(James et al. 2013)

Outliers

A subset of values that violate the assumed sampling distribution of Y.



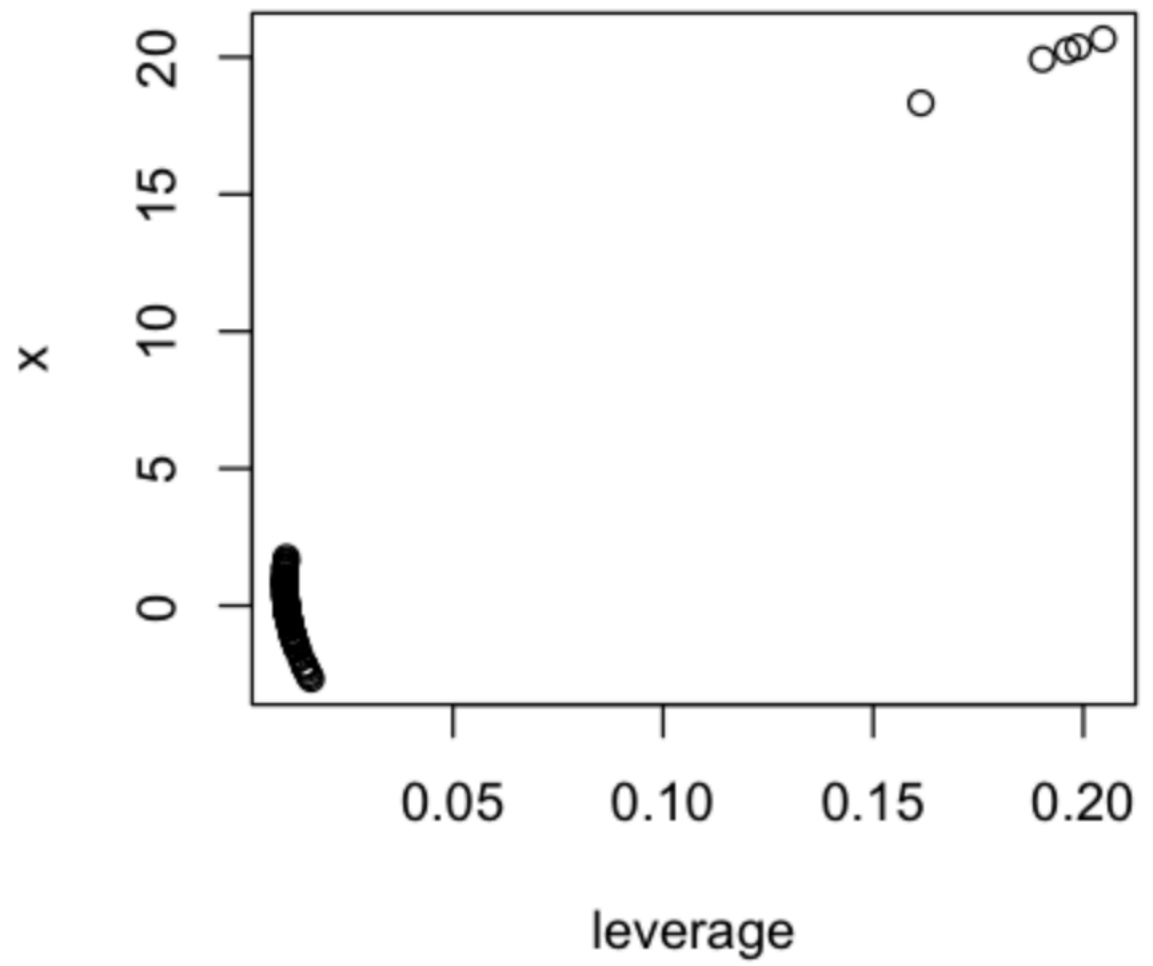
(James et al. 2013)

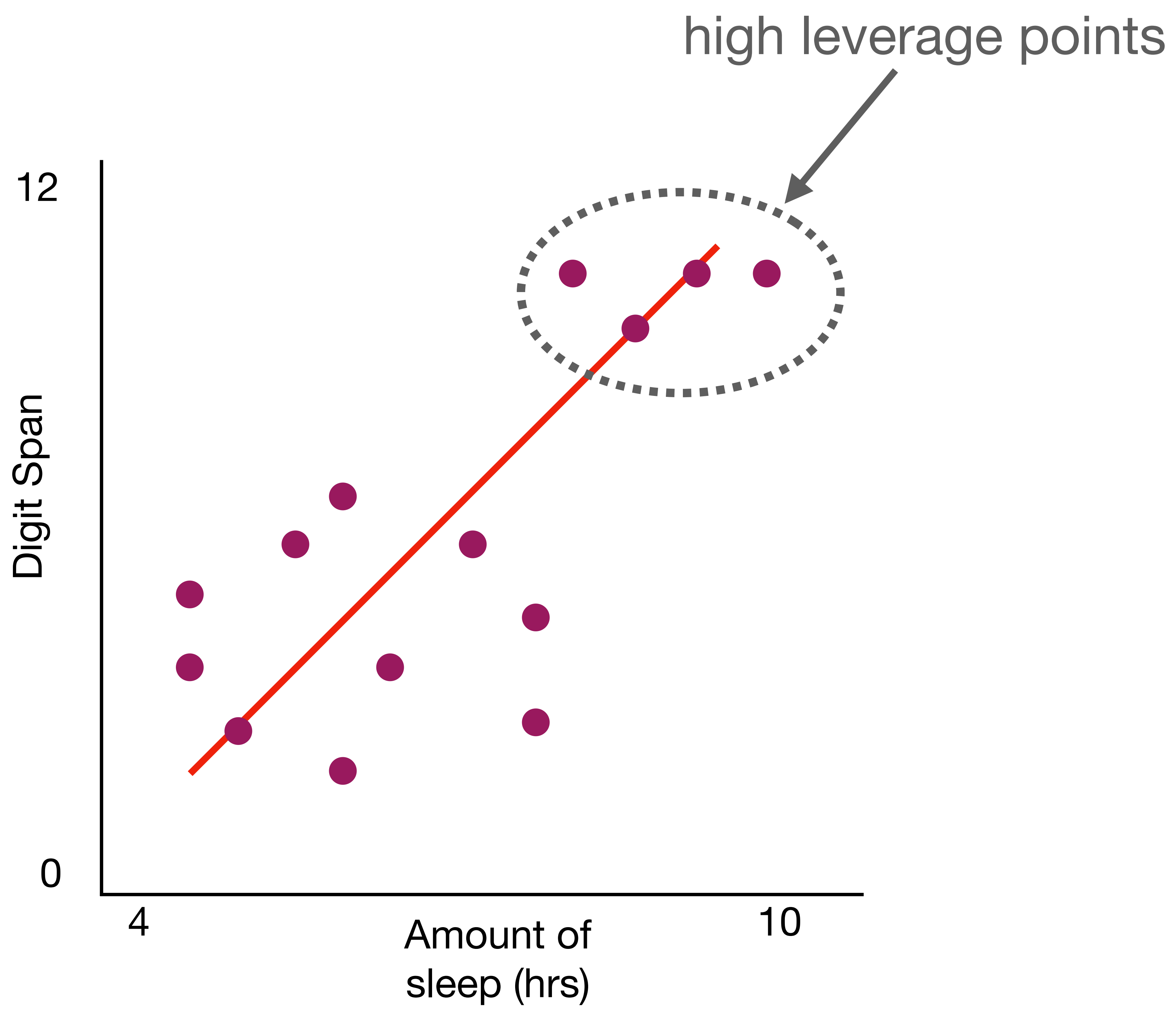
High leverage points

A subset of values in X that bias the outcome of $\hat{f}(X)$.

Leverage Statistic (h)

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$$





(James et al. 2013)

Collinearity

A correlation between two or more predictor variables in X.

Strong case: $Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$

Set: $X_1 = X_2 \rightarrow \rho(X_1, X_2) = 1$

Then:
$$
\begin{aligned}
Y &= \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 \\
&= \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_1 \\
&= \hat{\beta}_0 + X_1 \underbrace{(\hat{\beta}_1 + \hat{\beta}_2)}_{\tilde{\beta}_1} \\
&= \hat{\beta}_0 + \tilde{\beta}_1 X_1
\end{aligned}
$$

Variance Inflation Factor (VIF)

$$VIF = \frac{1}{1 - r^2_{X_j|X_{-j}}}$$

The closer to 1 that $r^2_{X_j|X_{-j}}$ gets, the more likely that the correlation is impacting $\hat{f}(X)$.

(James et al. 2013)

Take home message

- Interpreting linear regression models depends critically on how they are setup and how closely the data aligns to the assumptions of the fitting routine.