Building Intuition around Common Statistical Learning Techniques

Will Daniels
Department of Applied Mathematics and Statistics
Colorado School of Mines

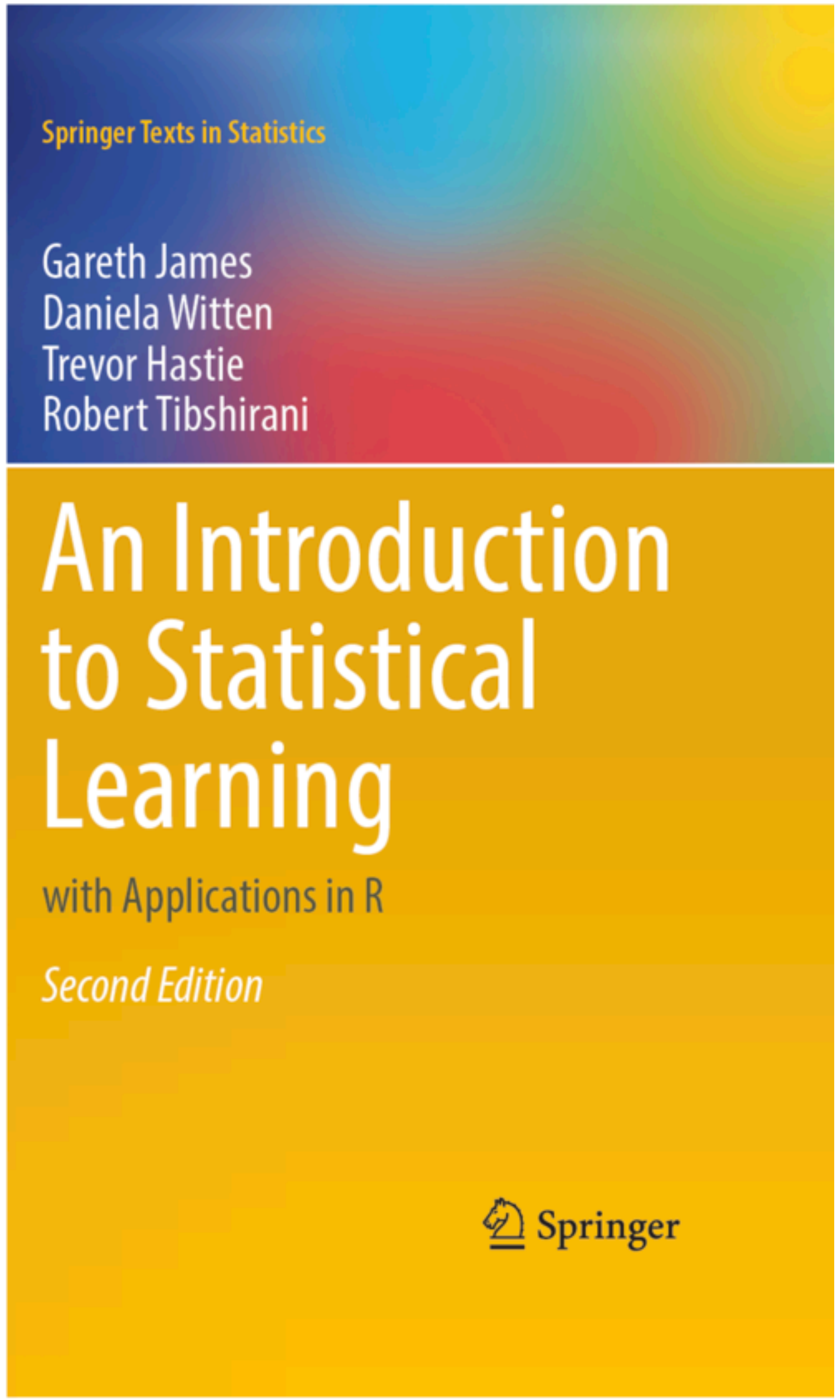
February 28, 2022

Three important principles

1. Statistical learning methods are useful in a wide range of disciplines
2. Statistical learning should not be viewed as a black box
3. While it is important to understand the strengths, weaknesses, and assumptions of each statistical learning method, it is not necessary to build them from scratch

Great reference text

- Free pdf at: https://www.statlearning.com/
- Most of the images in this talk taken from ISLR



Agenda

- Introduction: what is statistical learning?
- Two common problems statistical learning can address
  - Regression techniques and their interpretation
  - Classification techniques and their interpretation
- How to evaluate a statistical learning model?
- Stat learning example: What are the drivers of fire season intensity in MSEA?
- R implementation

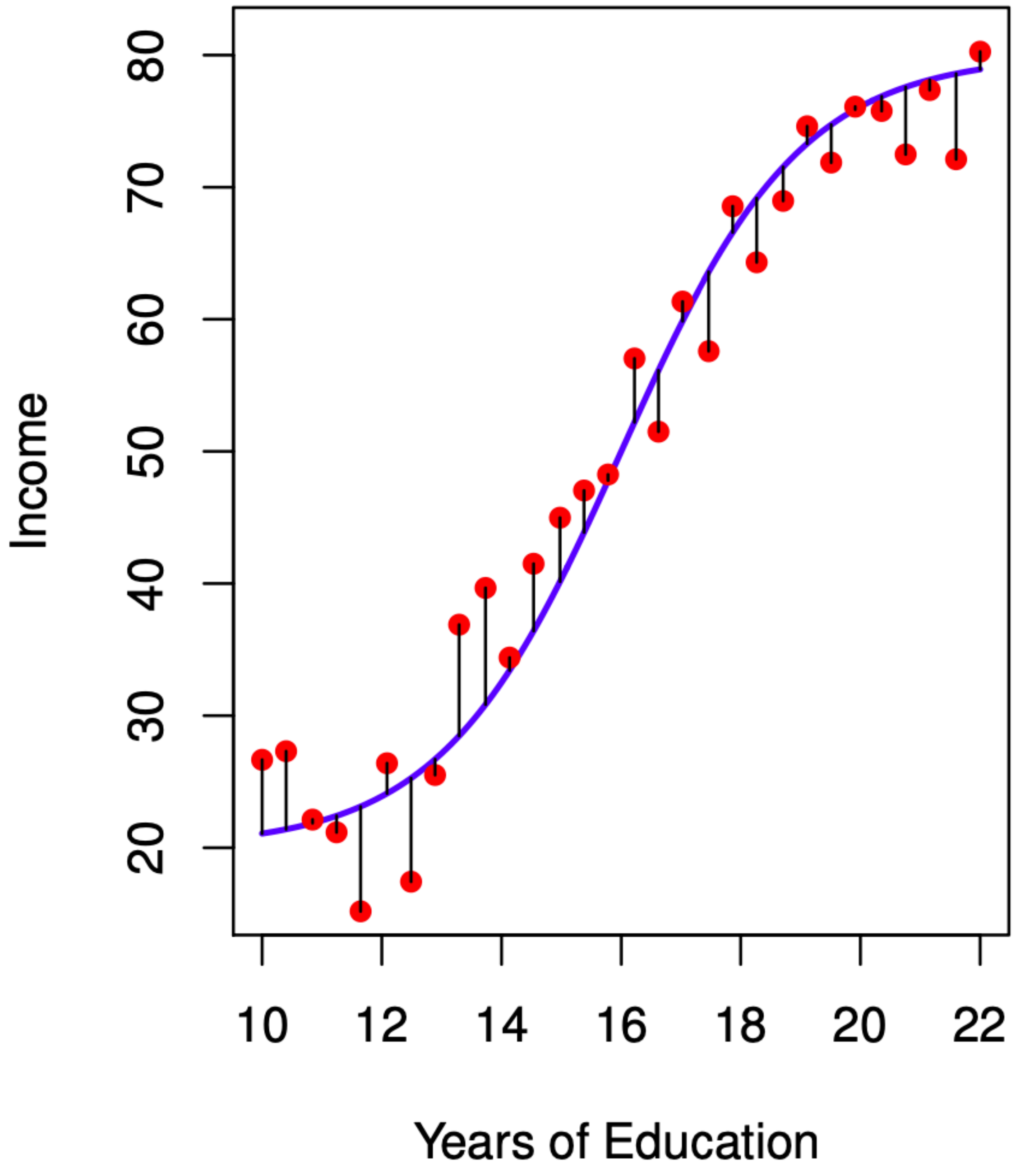
What is statistical learning?

What is statistical learning?

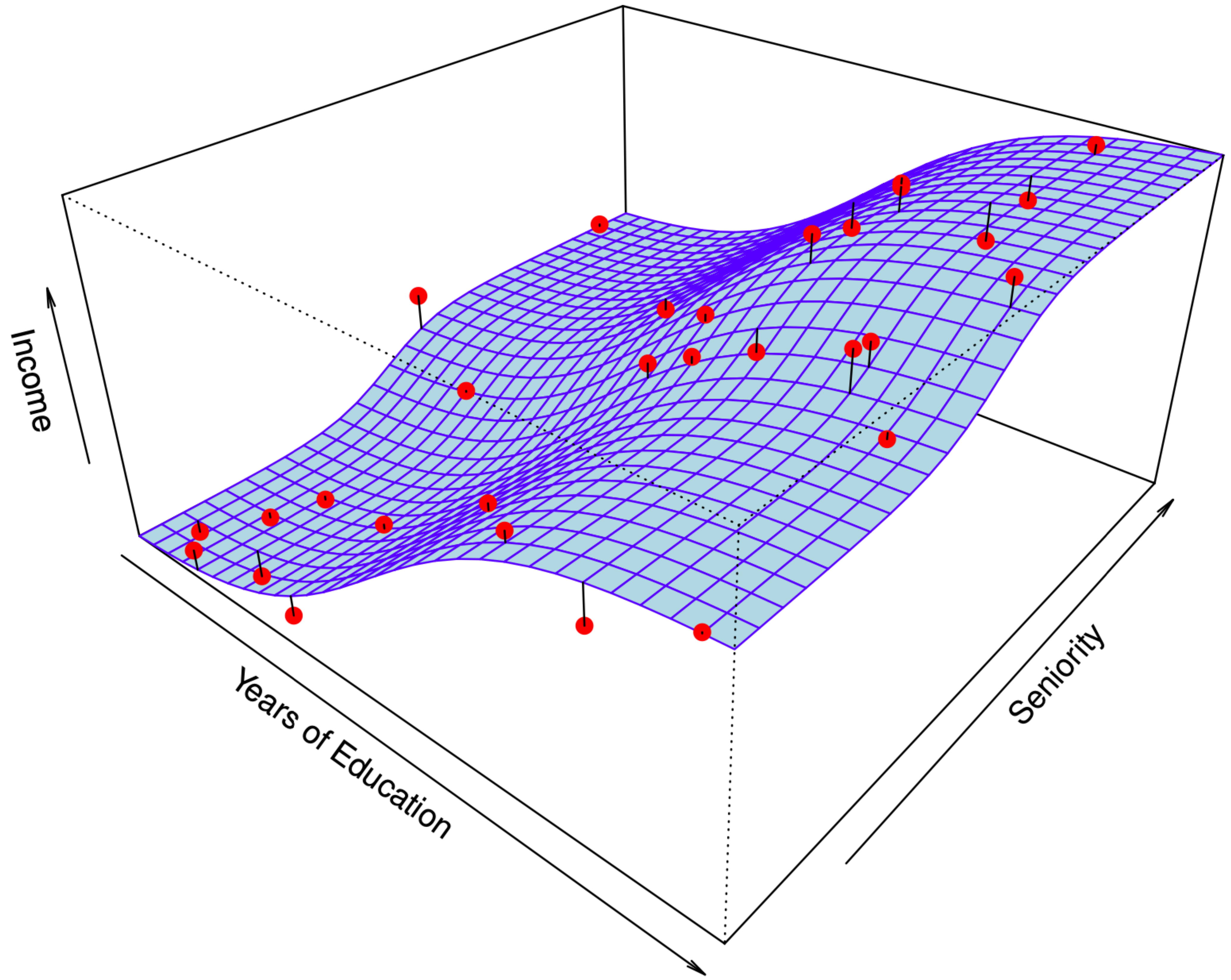
- Methods to estimate the relationship between variables (i.e., data)
- Given:
    - some response variable, Y
    - p different predictor variables, $X = (X_1, X_2, \ldots, X_p)$
- We assume a general relationship: $Y = f(X) + \epsilon$
    - f is some fixed but unknown function
    - ϵ is a random error term (usually mean zero)
- Statistical learning attempt to estimate the true relationship, f, with some approximation, $\hat{f}$

Example is linear regression: $Y = \beta_0 + \beta_1 X + \epsilon$

What is statistical learning?

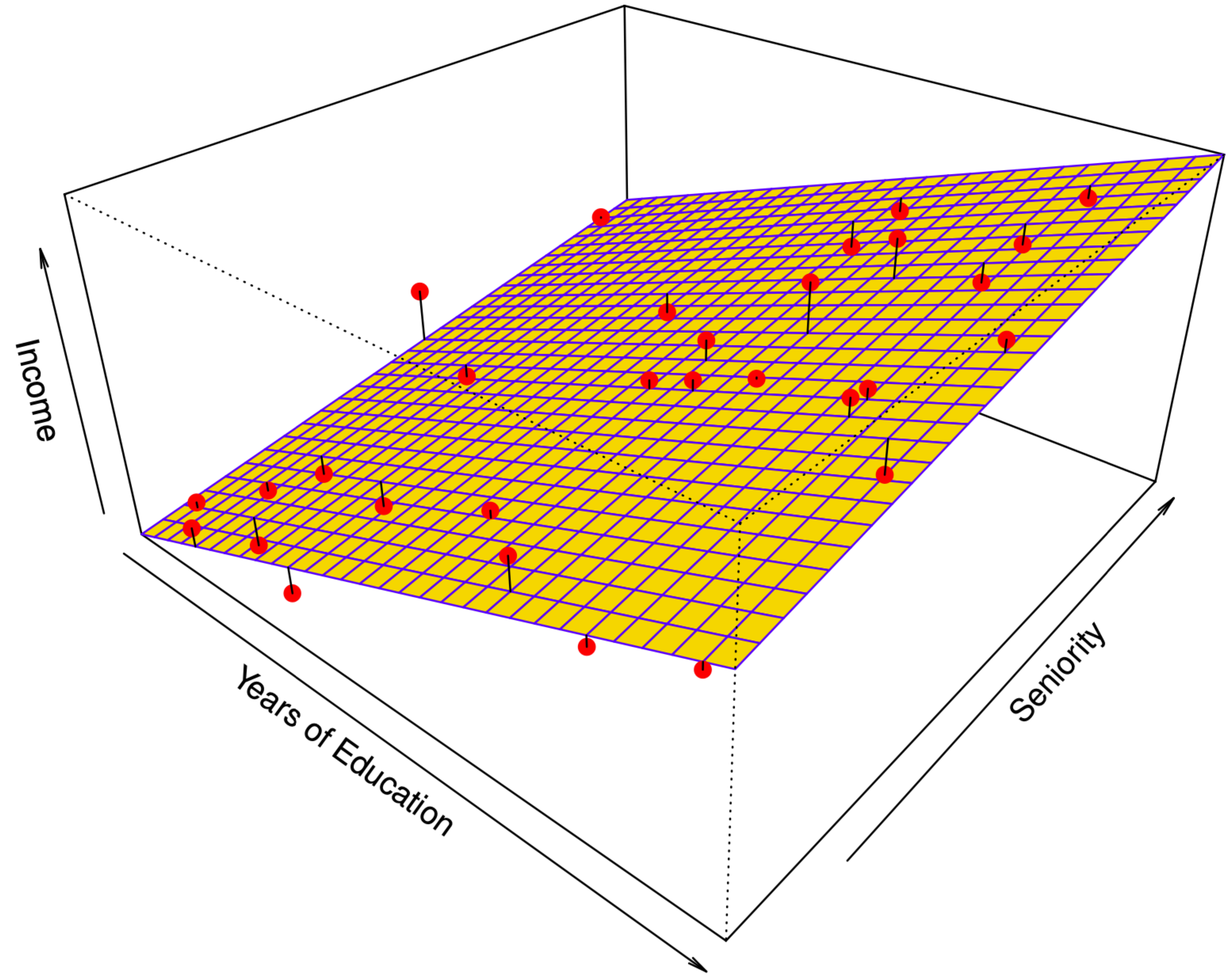


$X = (X_1)$

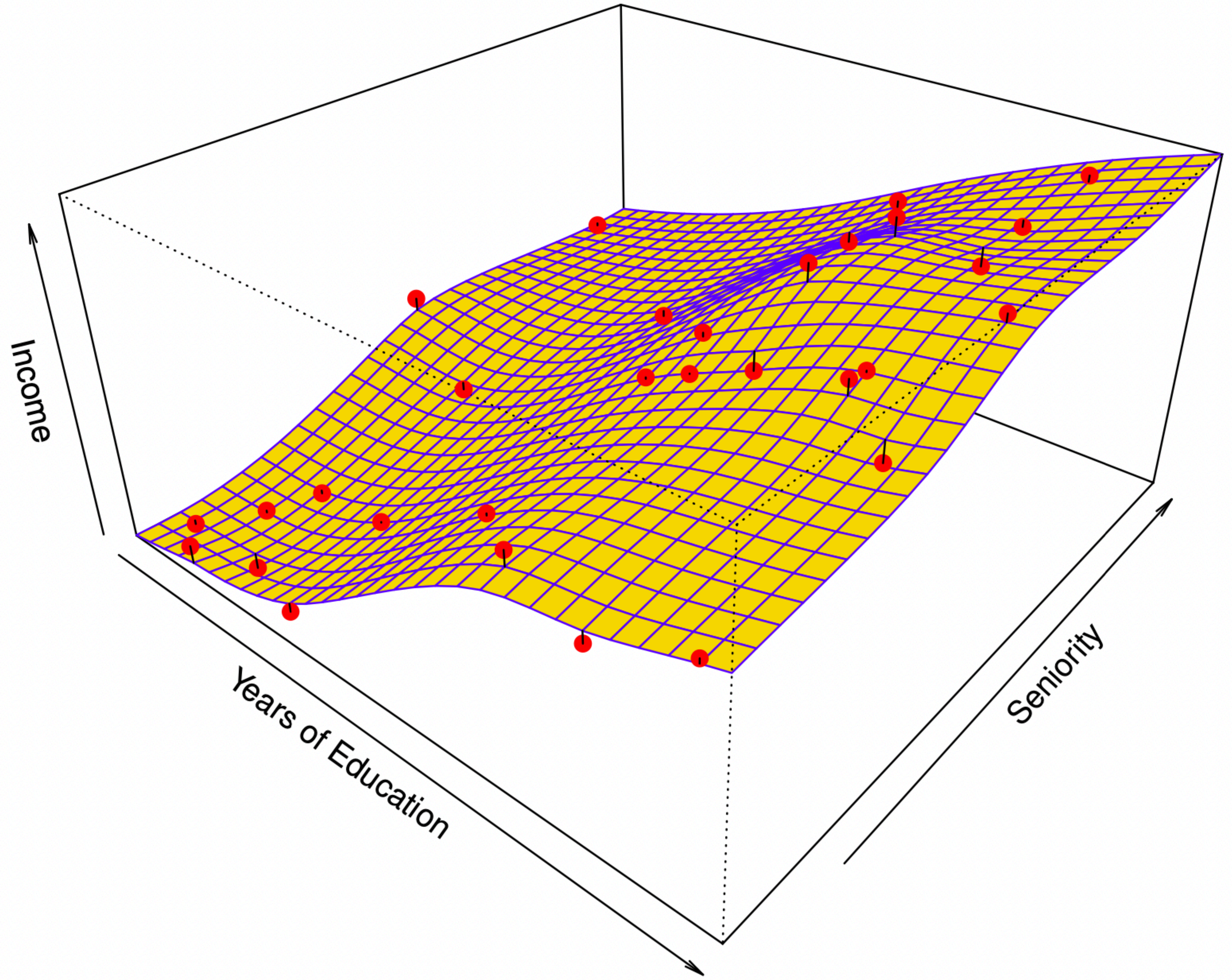


$X = (X_1, X_2)$

What is statistical learning?



$\hat{f}$ is least squares fit

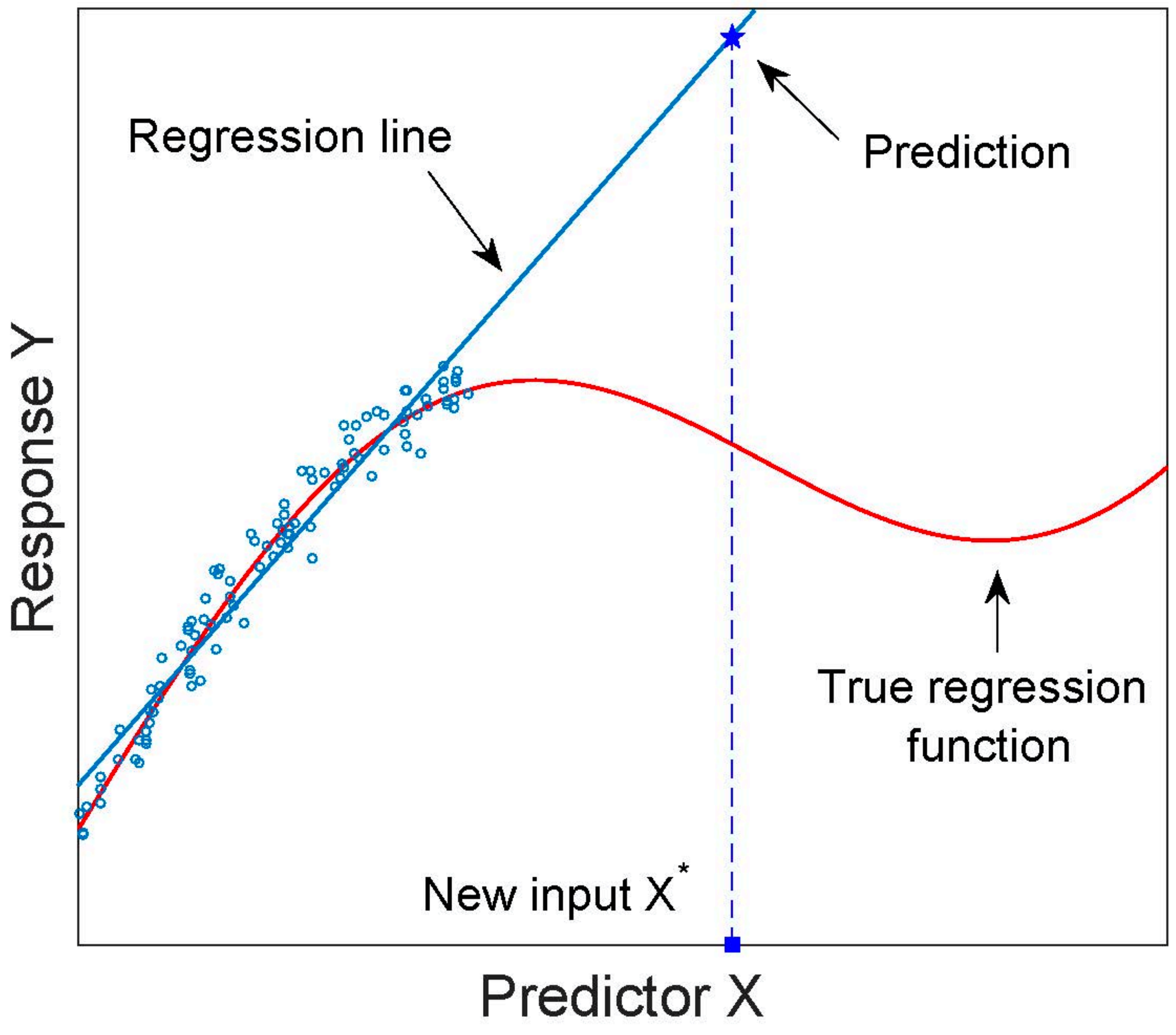


$\hat{f}$ is smooth thin-plate spline

Why estimate f?

1. Prediction

- Often X values are easy to obtain, but Y values require some effort to measure
- Since ϵ is often assumed to be mean zero, we can make predictions of Y using $\hat{f}$
- Predictions: $\hat{Y} = \hat{f}(X)$



Note: be careful with extrapolation!

Why estimate f?

2. Inference

- Want to better understand the association between Y and $X = (X_1, X_2, \ldots, X_p)$
  - Which of the $X_1, X_2, \ldots, X_p$ have an important association with Y?
  - What is the relative important of each $X_1, X_2, \ldots, X_p$ in explaining Y?
  - What is the form of the relationship? Linear? Non-linear?

When to use statistical learning

- Great at picking out relationships from data, but only when you have enough data
  - **Parametric models:** require less data because you specify a general form of the model (f)

    e.g. linear regression: $Y = \beta_0 + \beta_1 X + \epsilon$

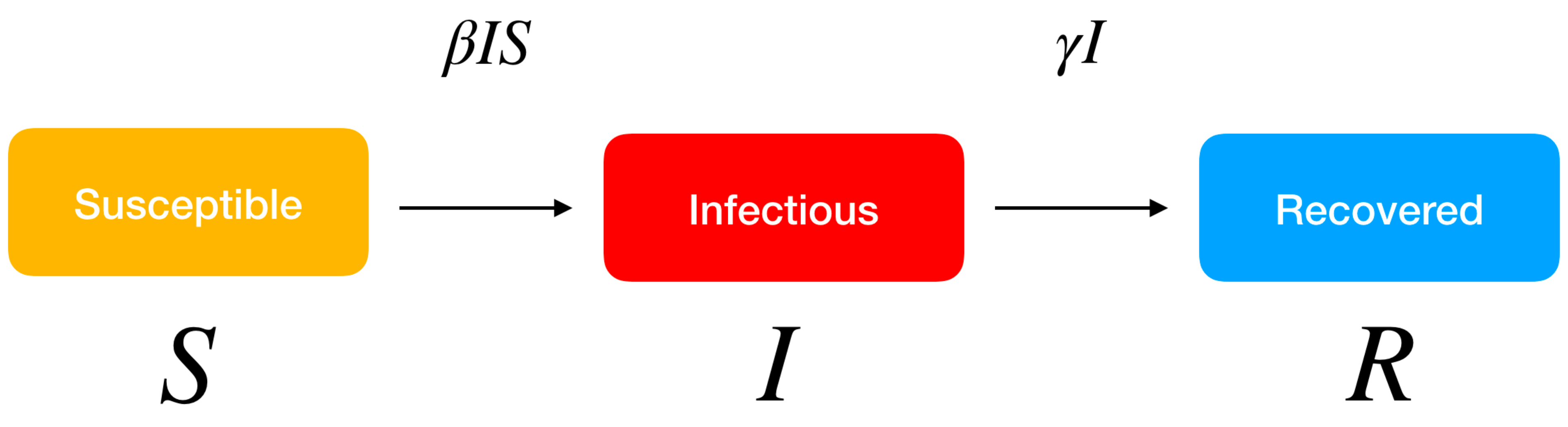
  - **Non-parametric models:** usually require more data because you don't specify a form of the model (f)

    e.g. smoothing spline: $\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int f''(t)^2 dt$

- Can be very flexible, interpretable, and accurate
- Usually come with some way of performing uncertainty quantification

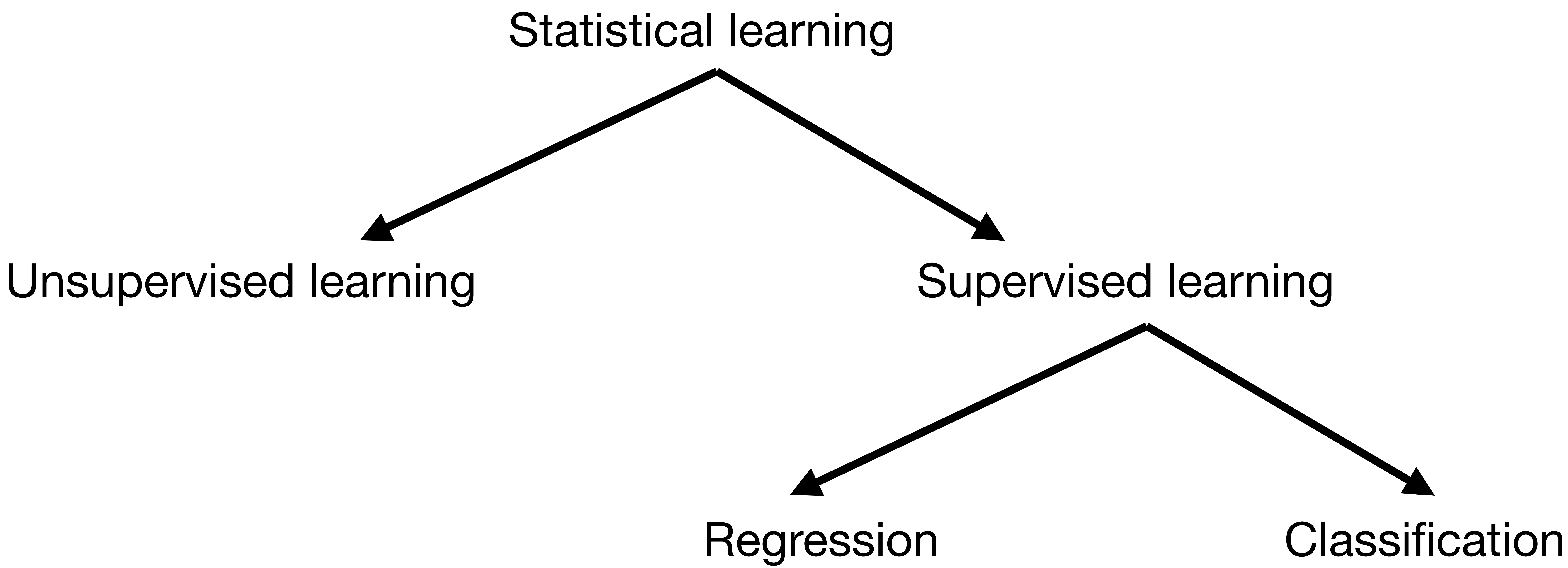
When to consider another model

- There is not enough data to properly train / estimate parameter values
- Example:
  - Modeling hospitalizations from Omicron.
  - Could use a mechanistic model instead (e.g., SIR ODE model)

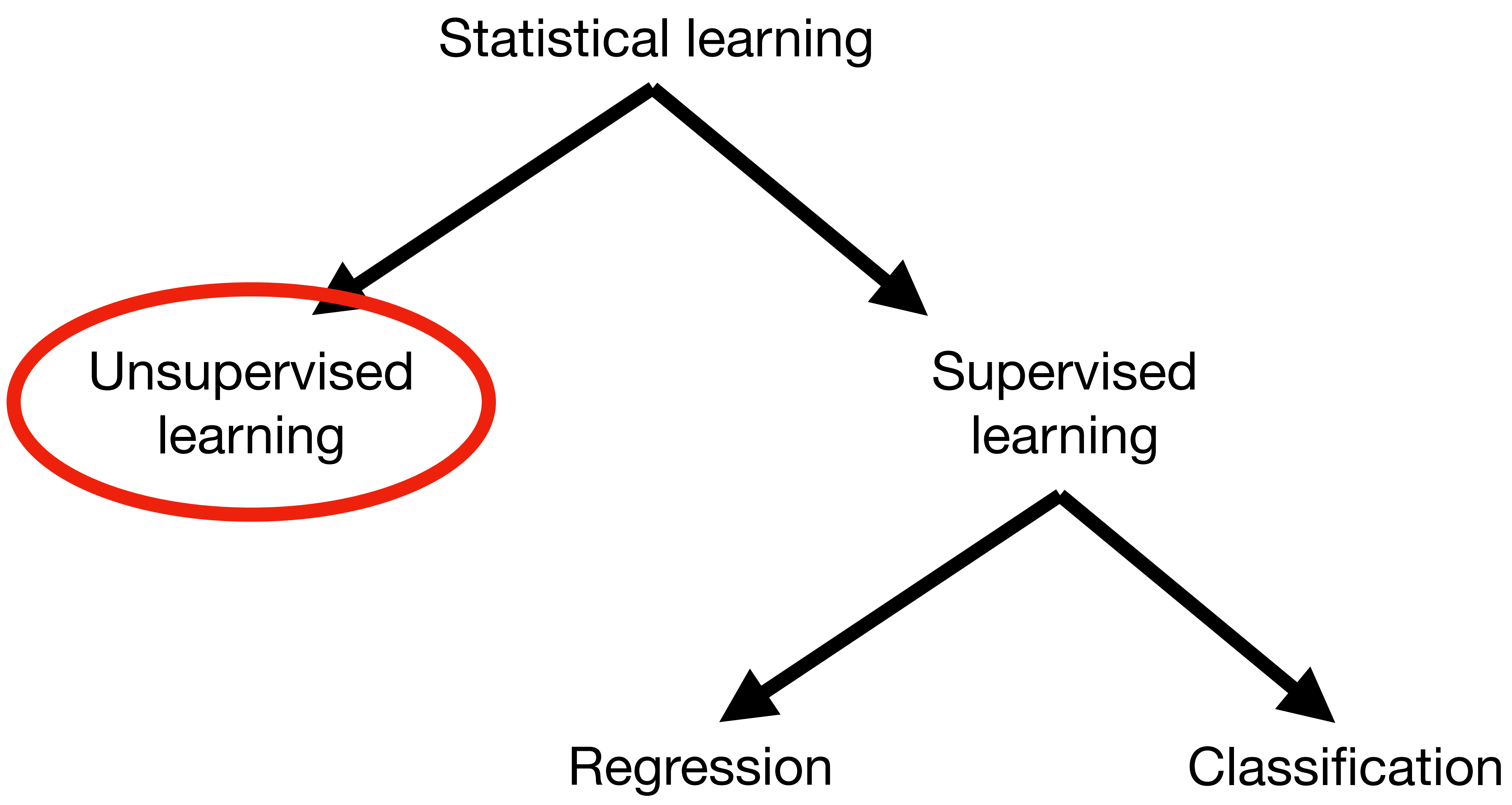


$$\begin{cases} \frac{dS}{dt} = -\frac{\beta I S}{N}, \\ \frac{dI}{dt} = \frac{\beta I S}{N} - \gamma I, \\ \frac{dR}{dt} = \gamma I, \end{cases}$$

Types of problems that statistical learning can address

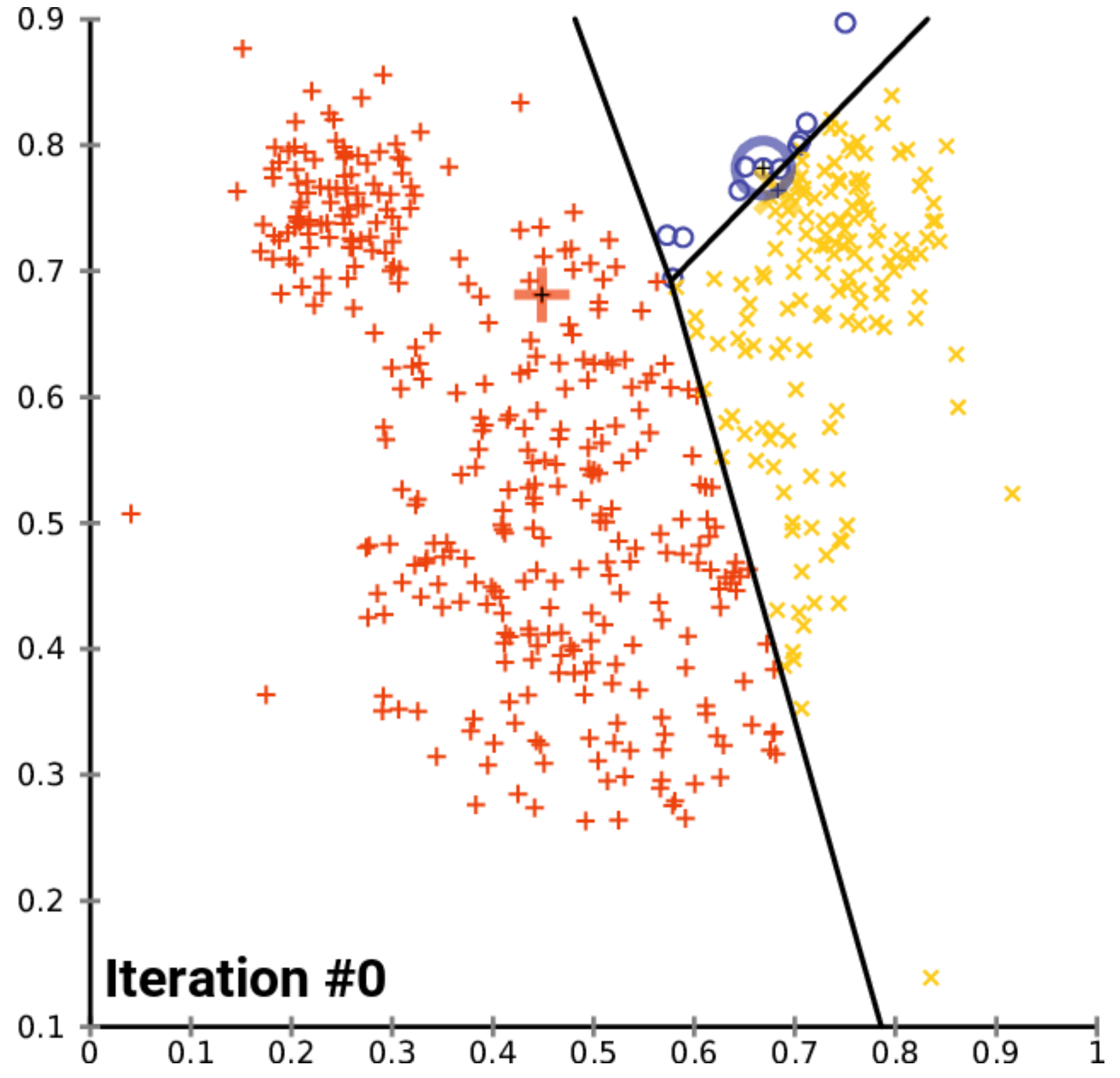


Types of problems that statistical learning can address

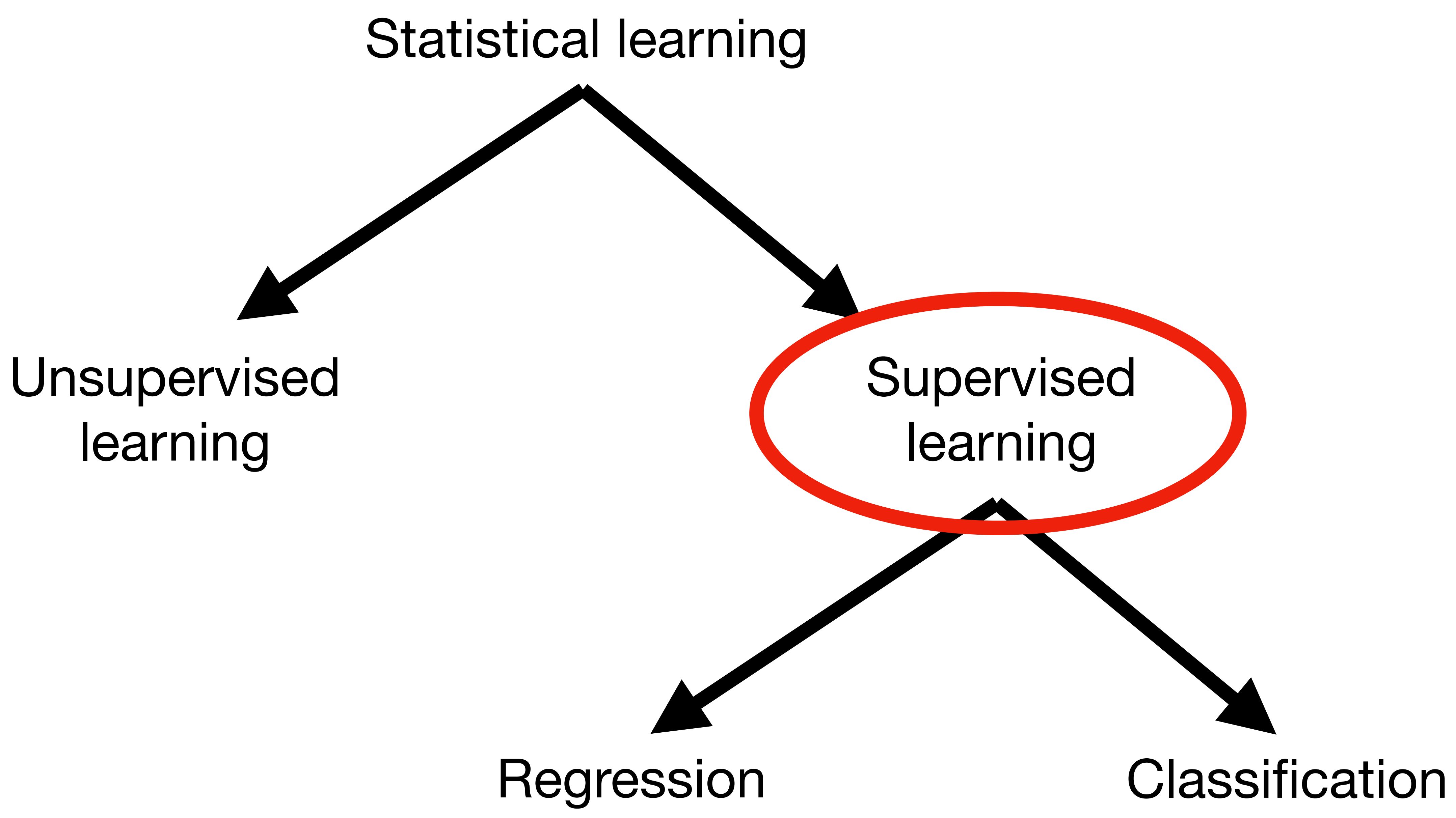


Unsupervised Learning
We observe measurements $X = (X_1, X_2, ..., X_p)$ but no associated response Y

Example: k-means clustering



Types of problems that statistical learning can address



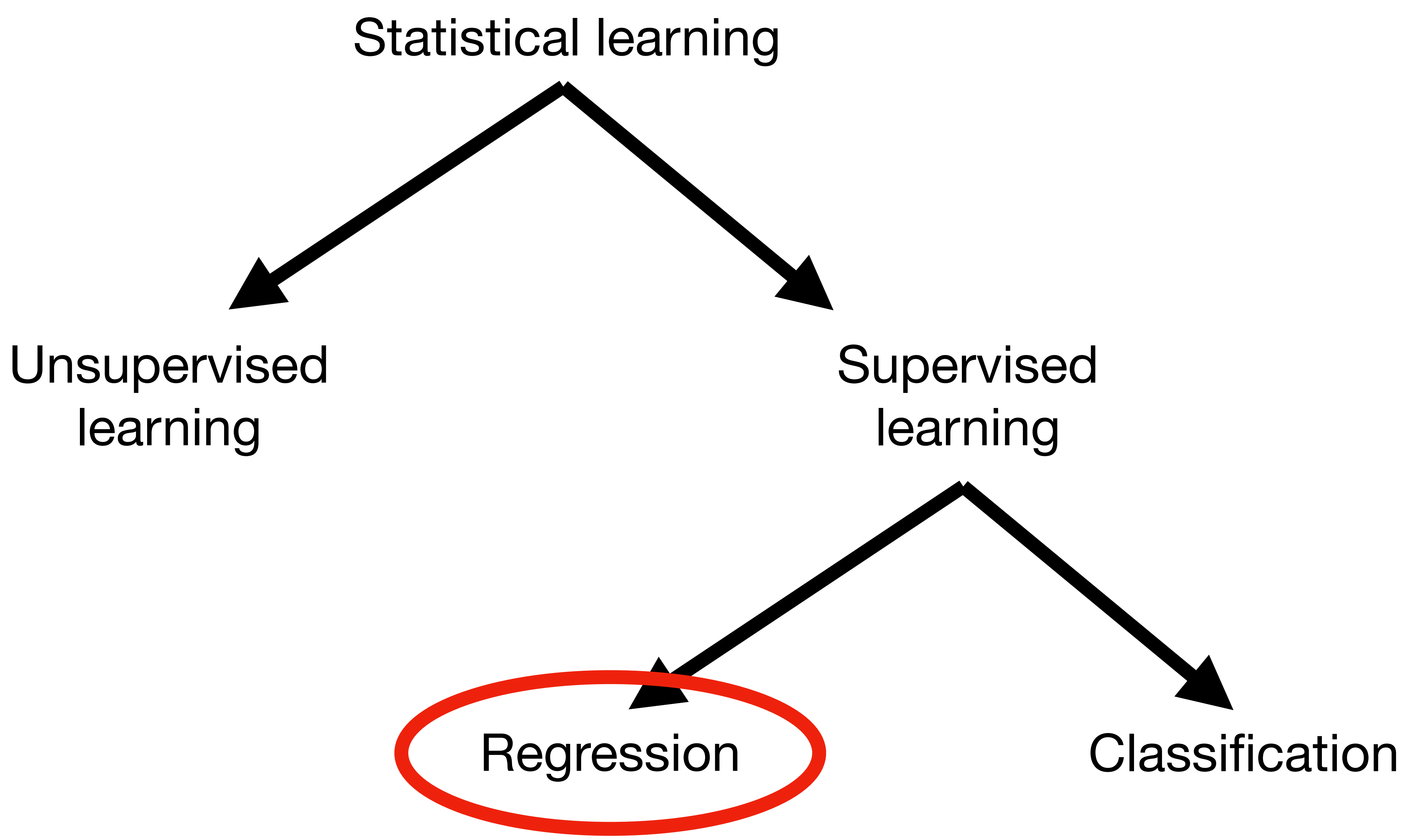
Supervised Learning

We observe measurements $X = (X_1, X_2, \ldots, X_p)$ and associated response Y

Can be divided into two problems based on the form of Y

- **Regression** - model a continuous response
- **Classification** - model a categorical response

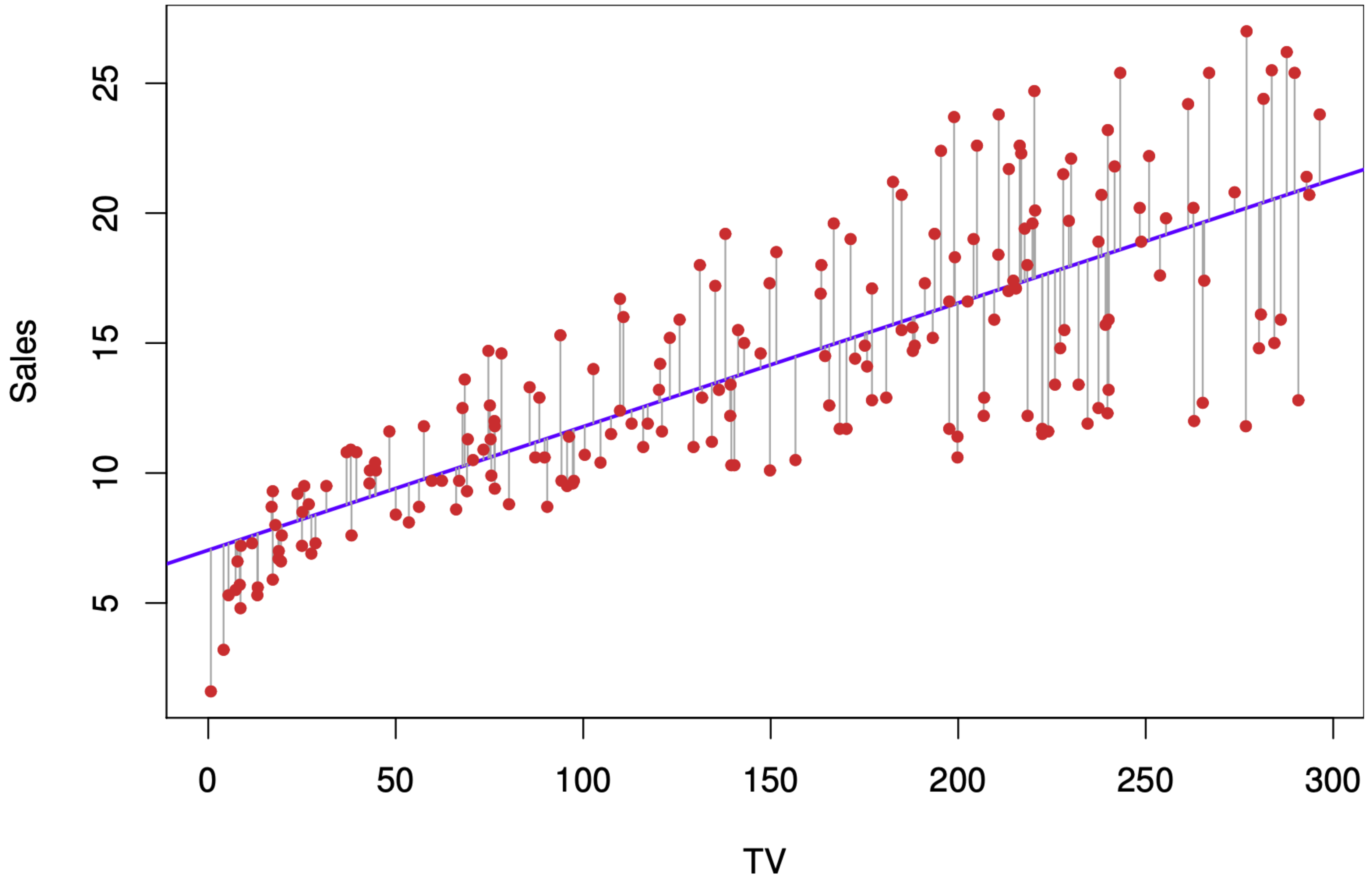
Types of problems that statistical learning can address



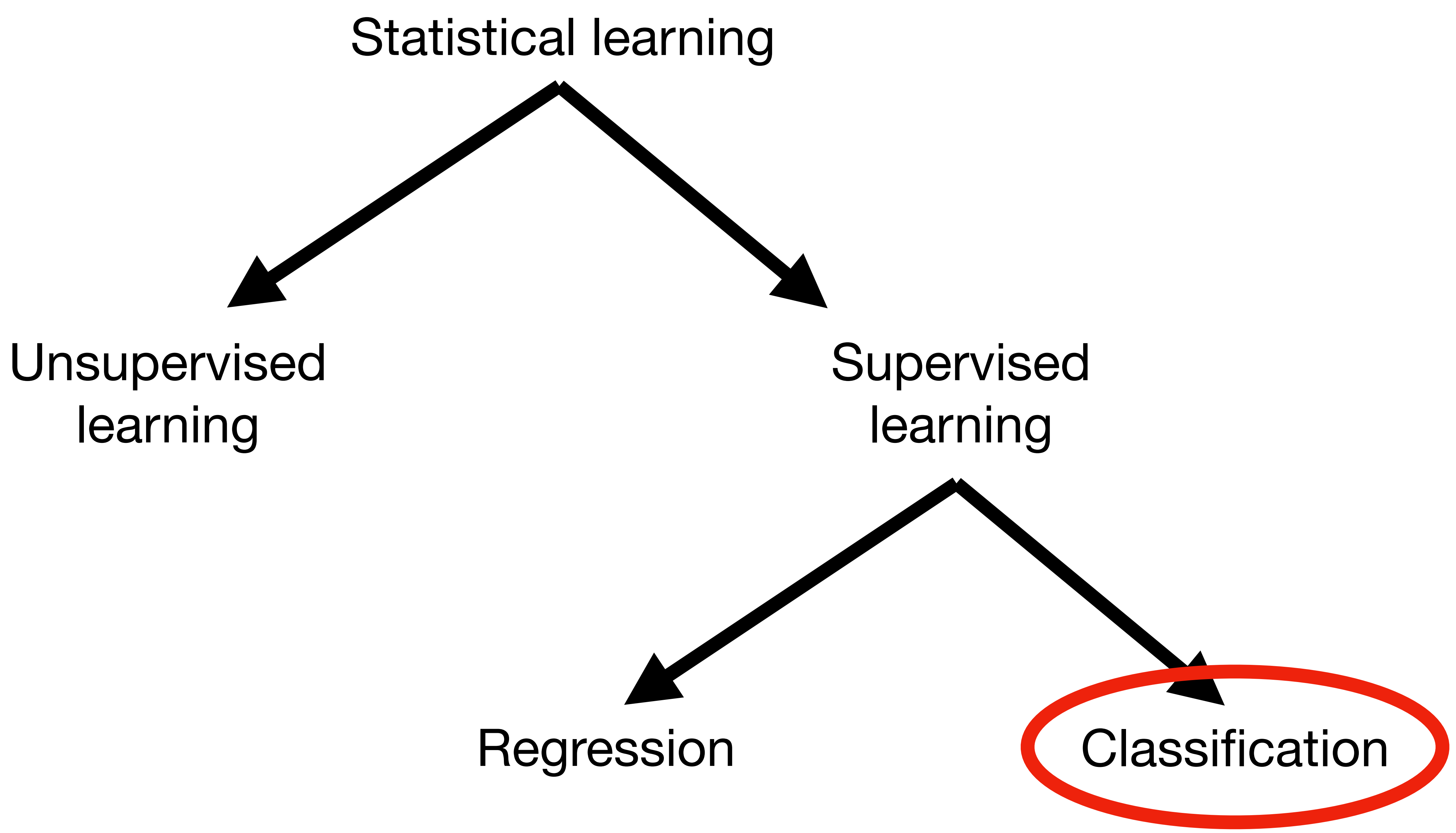
Regression

We observe measurements $X = (X_1, X_2, ..., X_p)$ and associated response Y that takes **continuous** values

Example: simple linear regression



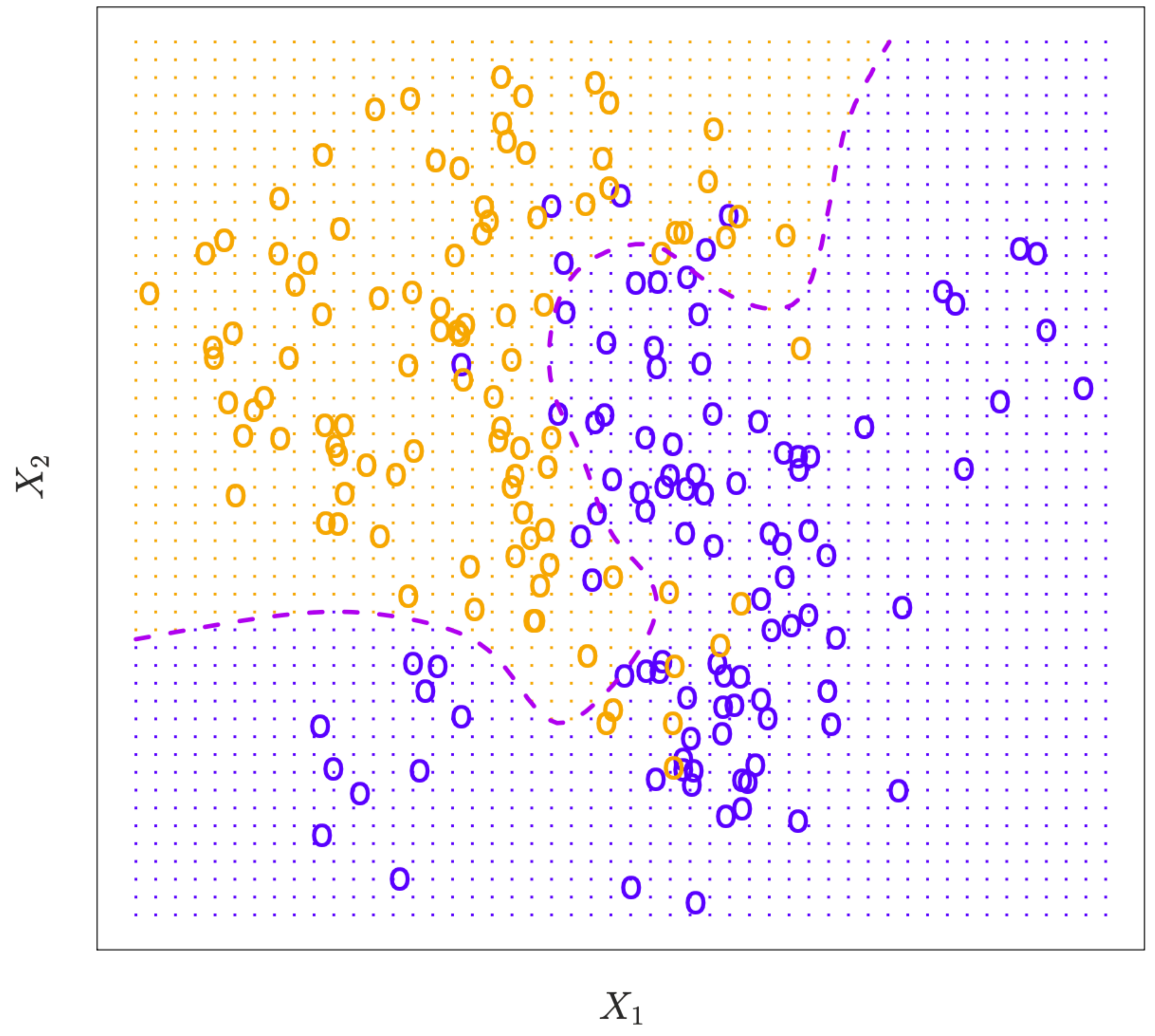
Types of problems that statistical learning can address



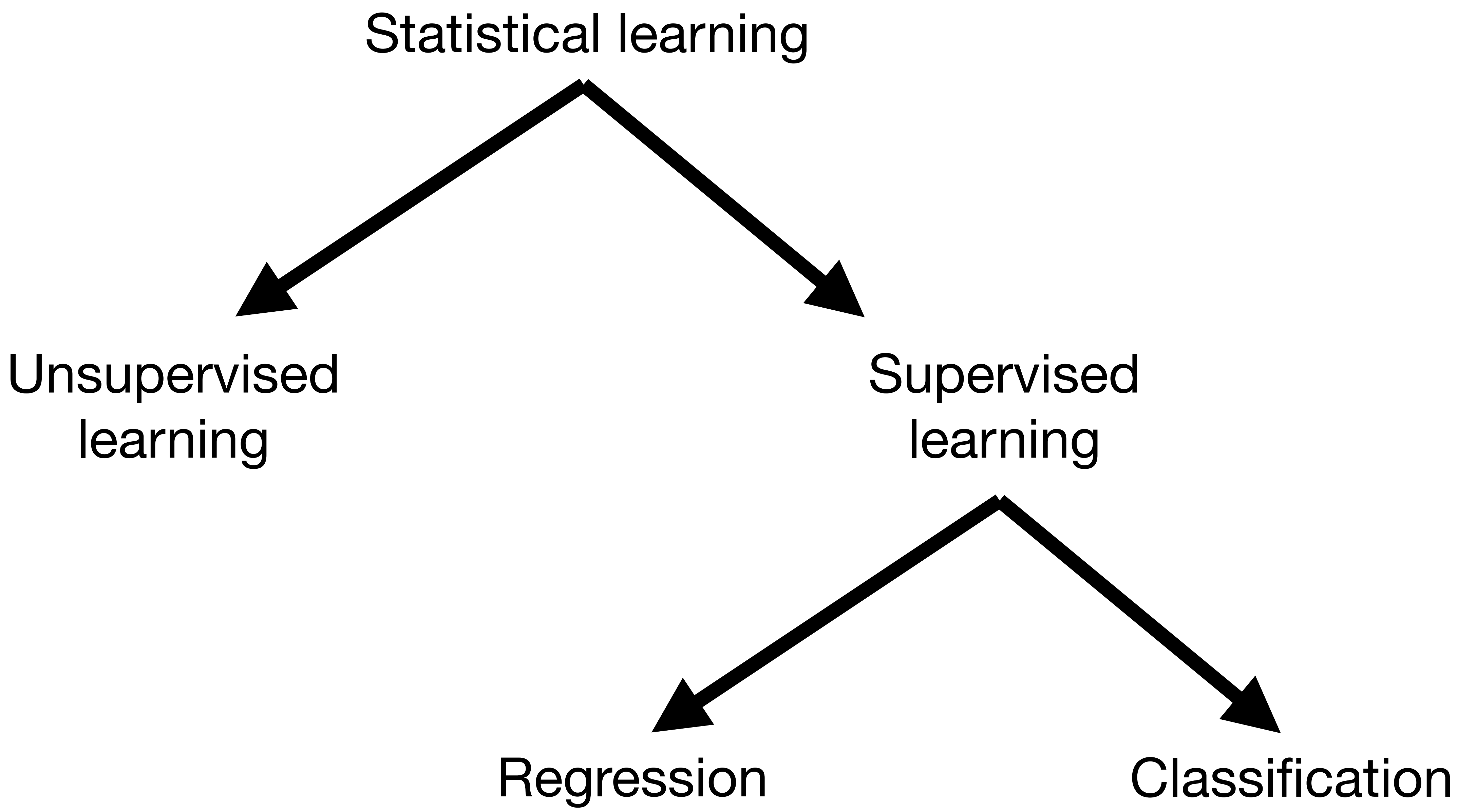
Classification

We observe measurements $X = (X_1, X_2, ..., X_p)$ and associated response Y that takes **categorical** values

Example: k-nearest neighbors



Regression

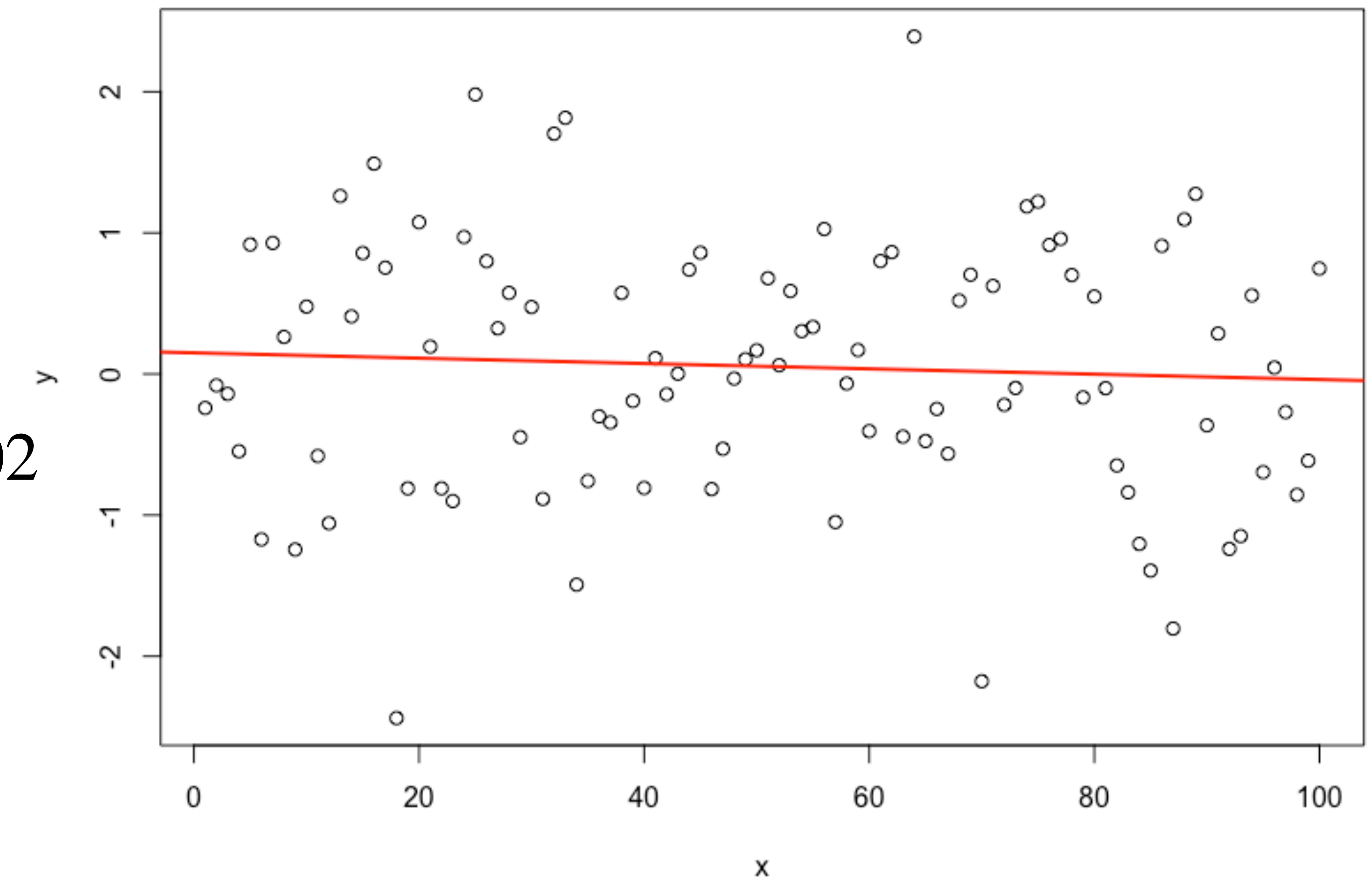

Statistical learning
Unsupervised learning
Supervised learning
Regression
Classification

Simple linear regression

- Assume a model of the form: $Y = \beta_0 + \beta_1 X + \epsilon$, where X is a single variable!
- Just have to estimate two parameters for prediction and inference: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
- Good for answering:
  - Is there a relationship between Y and X? How strong is this relationship? Is it linear?
  - Can we make accurate predictions of Y using X?

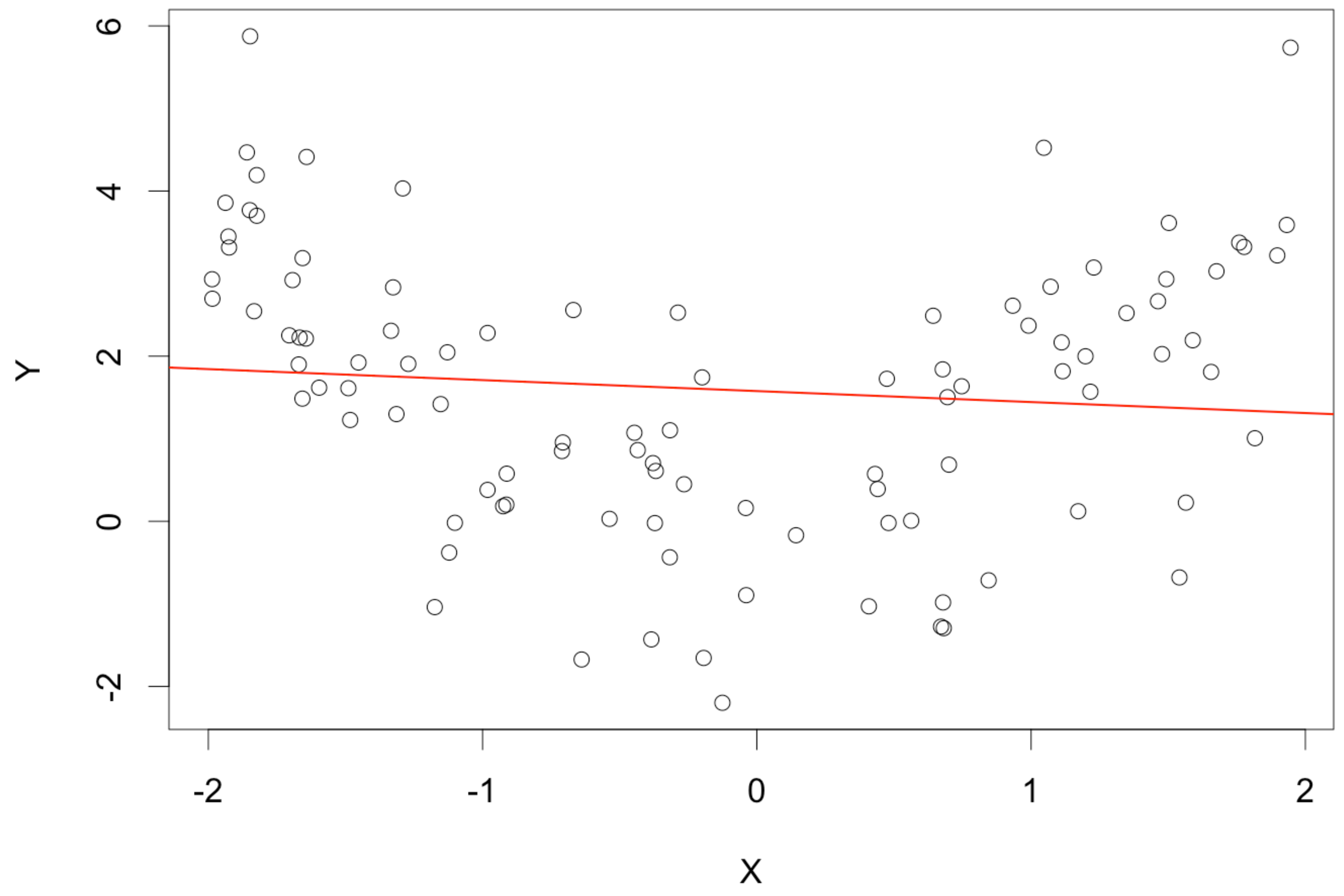
$\hat{\beta}_0 = 0.15$

$\hat{\beta}_1 = -0.002$



Simple linear regression

- Assume a model of the form: $Y = \beta_0 + \beta_1 X + \epsilon$
- Assumptions:
    1. There is a linear relationship between Y and X



Simple linear regression

Residuals = observations - predictions
$= Y - \hat{Y}$

- Assume a model of the form: $Y = \beta_0 + \beta_1 X + \epsilon$
- Assumptions:

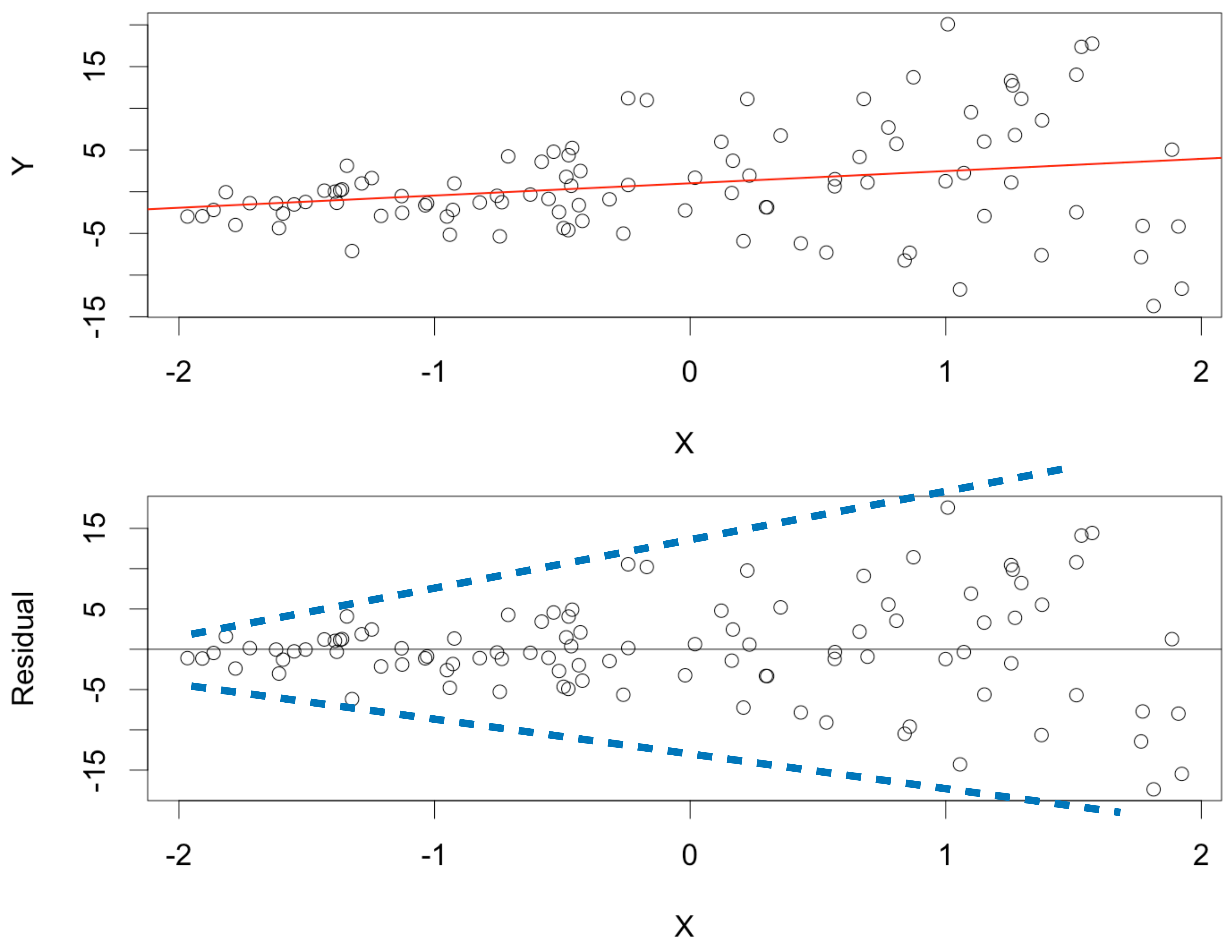
1. There is a linear relationship between Y and X
2. Independent residuals. How was the data collected?

Simple linear regression

Residuals = observations - predictions
$= Y - \hat{Y}$

- Assume a model of the form: $Y = \beta_0 + \beta_1 X + \epsilon$
- Assumptions:

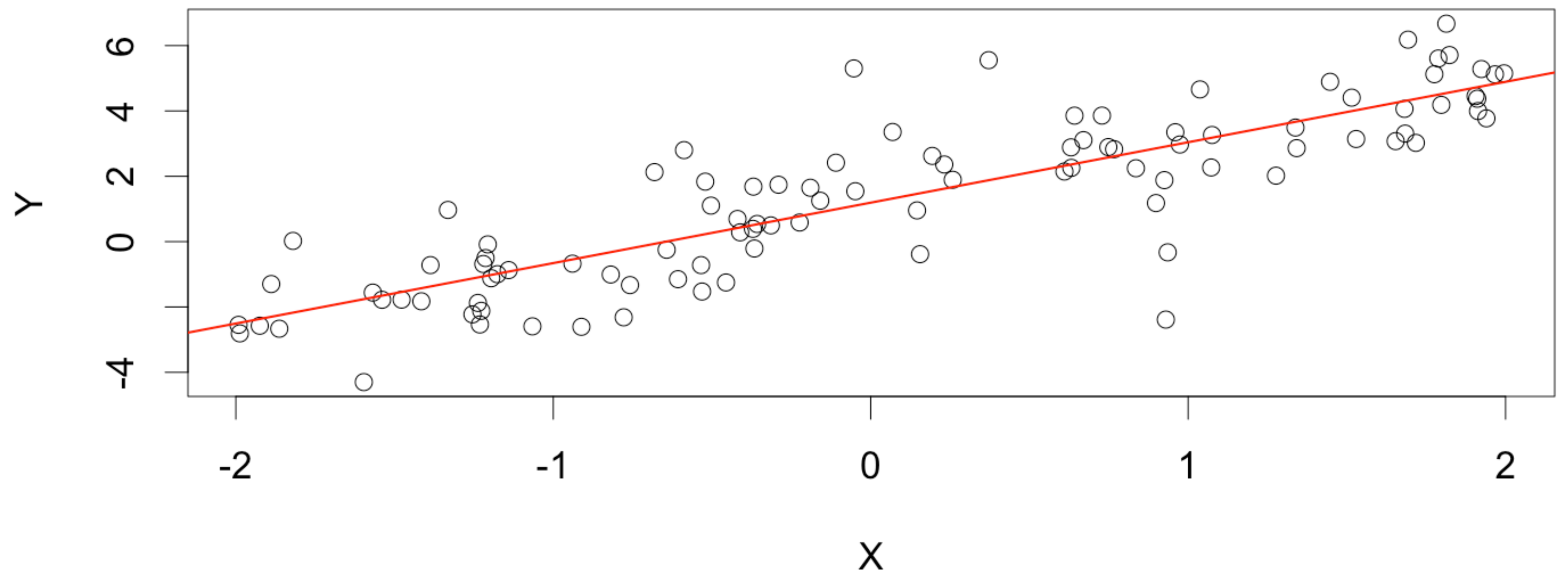
1. There is a linear relationship between Y and X
2. Independent residuals.
3. Residuals have constant variance.

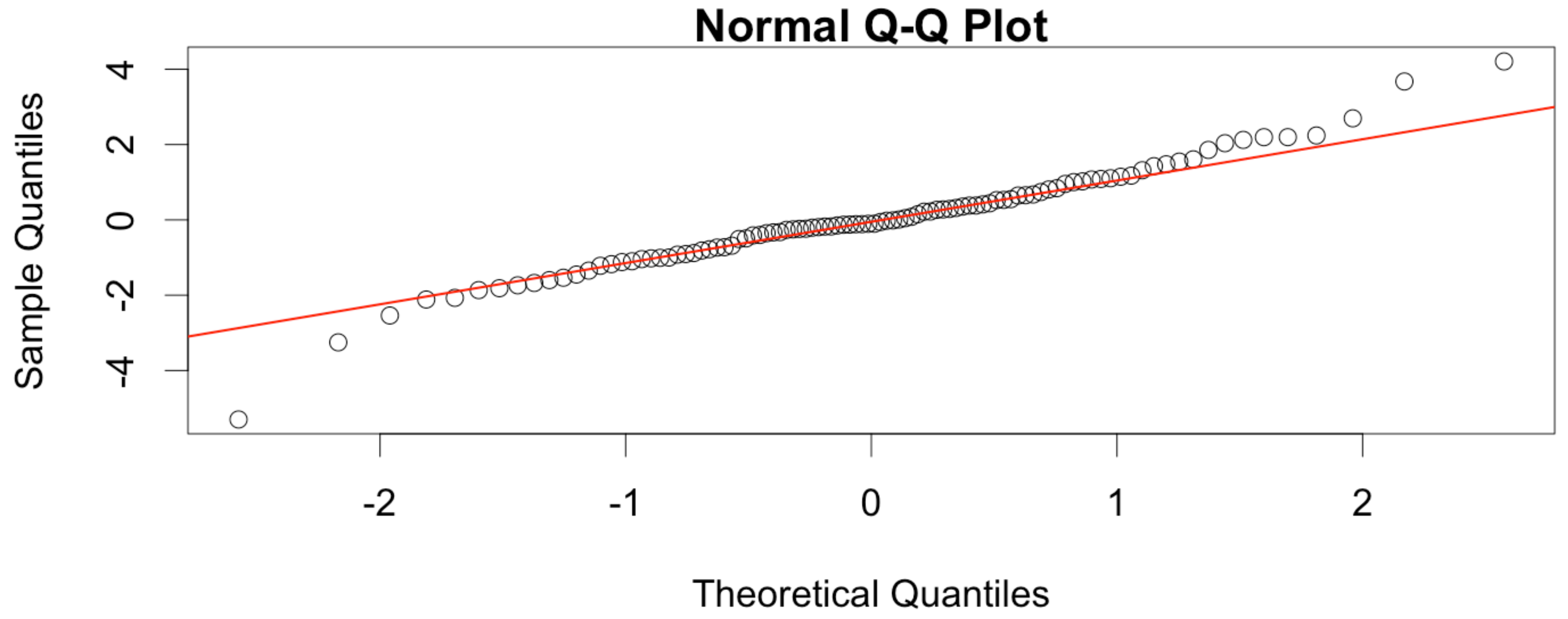


Simple linear regression

Residuals = observations - predictions
$= Y - \hat{Y}$

- Assume a model of the form: $Y = \beta_0 + \beta_1 X + \epsilon$
- Assumptions:
  1. There is a linear relationship between Y and X
  2. Independent residuals.
  3. Residuals have constant variance.
  4. Residuals are normally distributed.





Simple linear regression

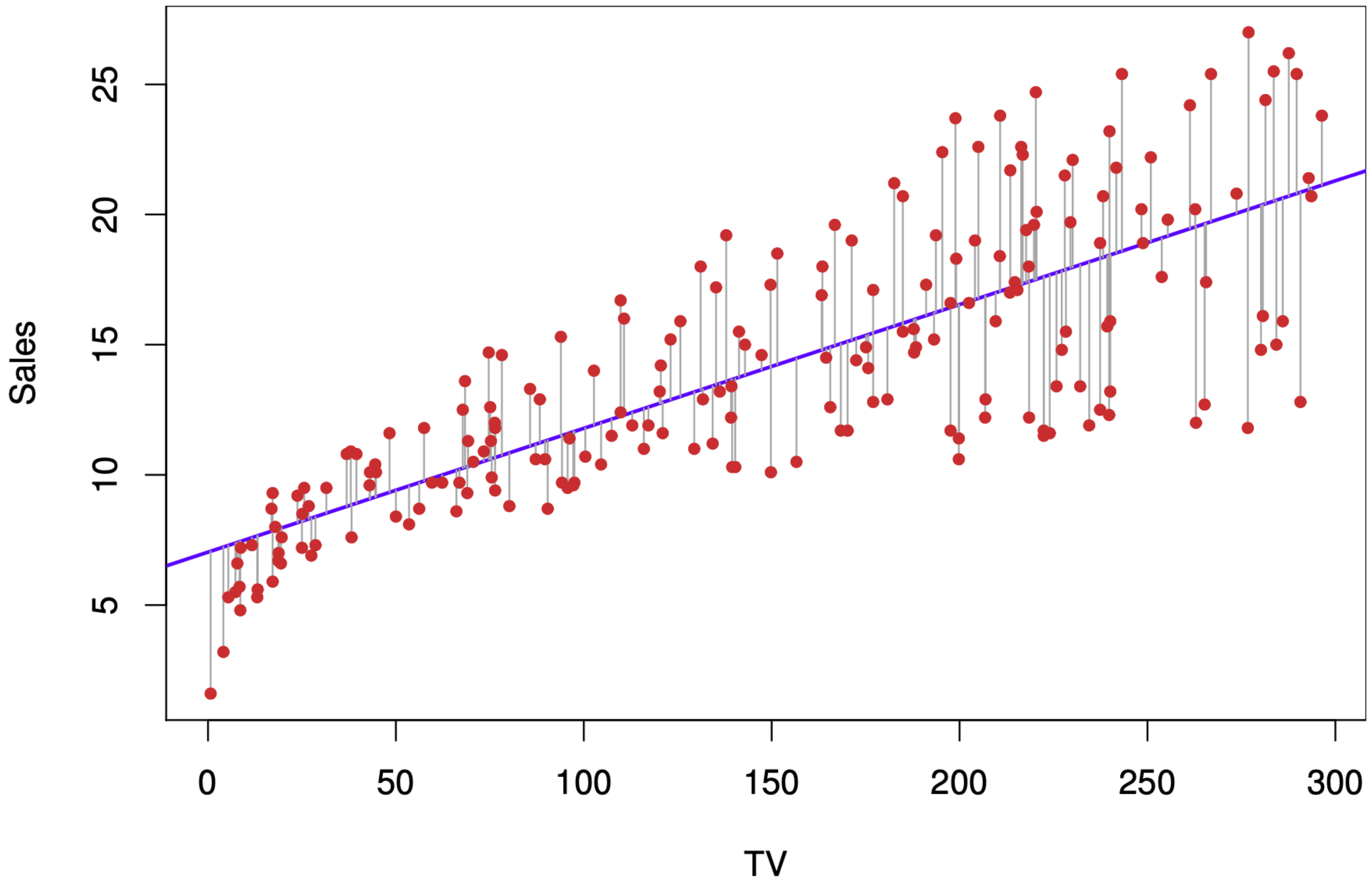
- Assume a model of the form: $Y = \beta_0 + \beta_1 X + \epsilon$

- Example: sales $= \beta_0 + \beta_1 \times$ TV

  - Is there a relationship between Y and X?

  - How strong is this relationship?

  - Is it linear?



Simple linear regression

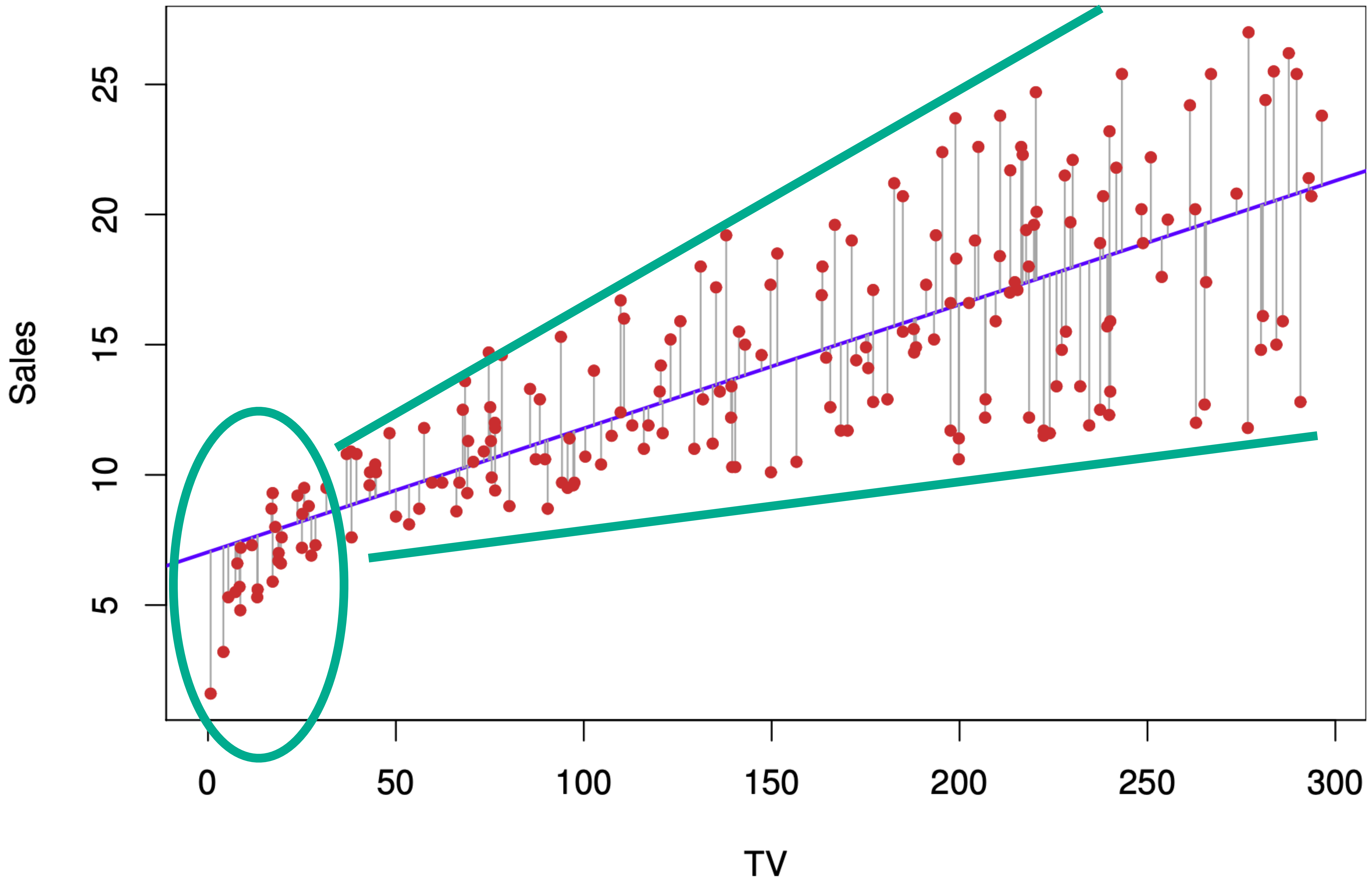
- Assume a model of the form: $Y = \beta_0 + \beta_1 X + \epsilon$

- Example: sales $= \beta_0 + \beta_1 \times$ TV

  - Is there a relationship between Y and X?

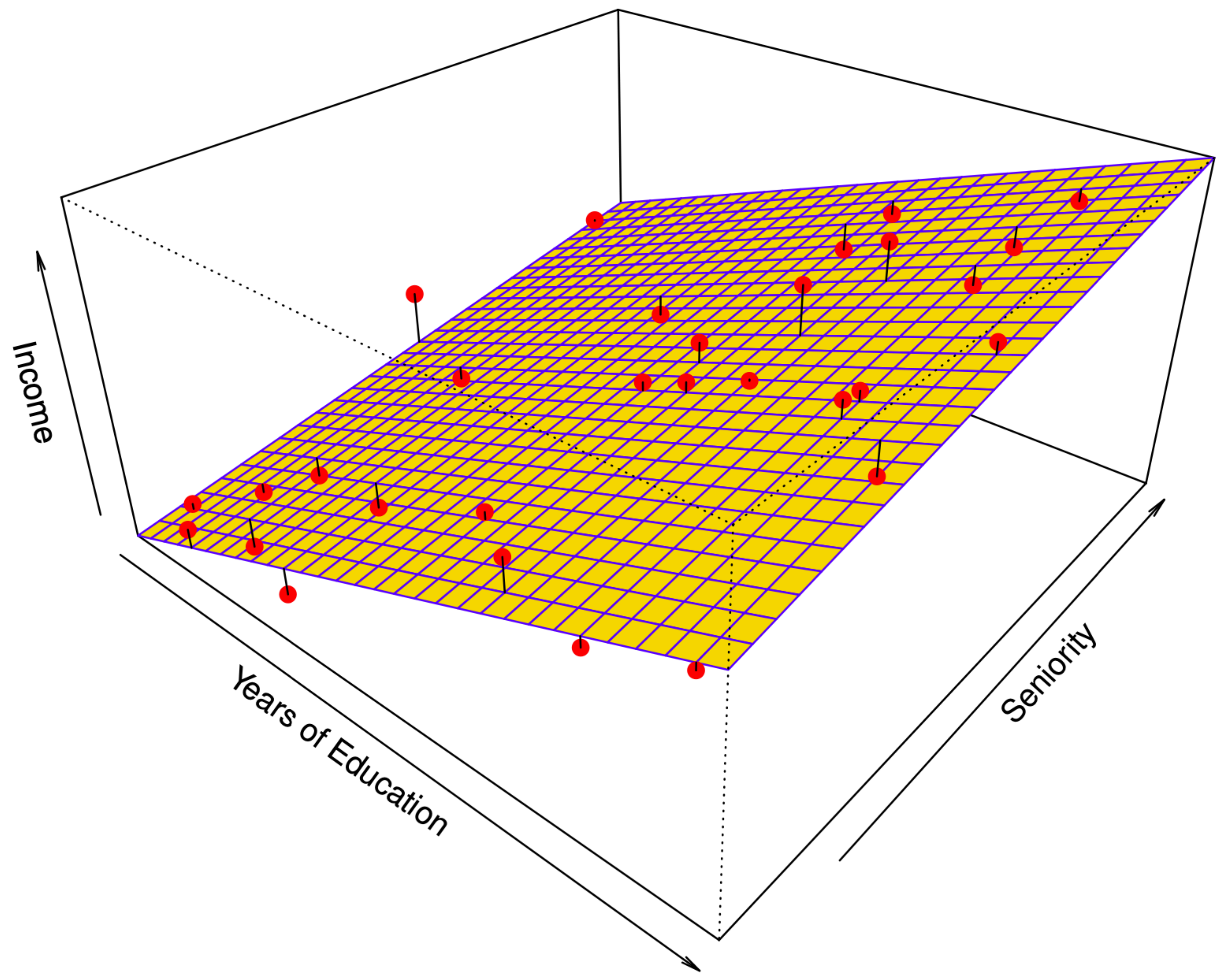
  - How strong is this relationship?

  - Is it linear?



Multiple linear regression

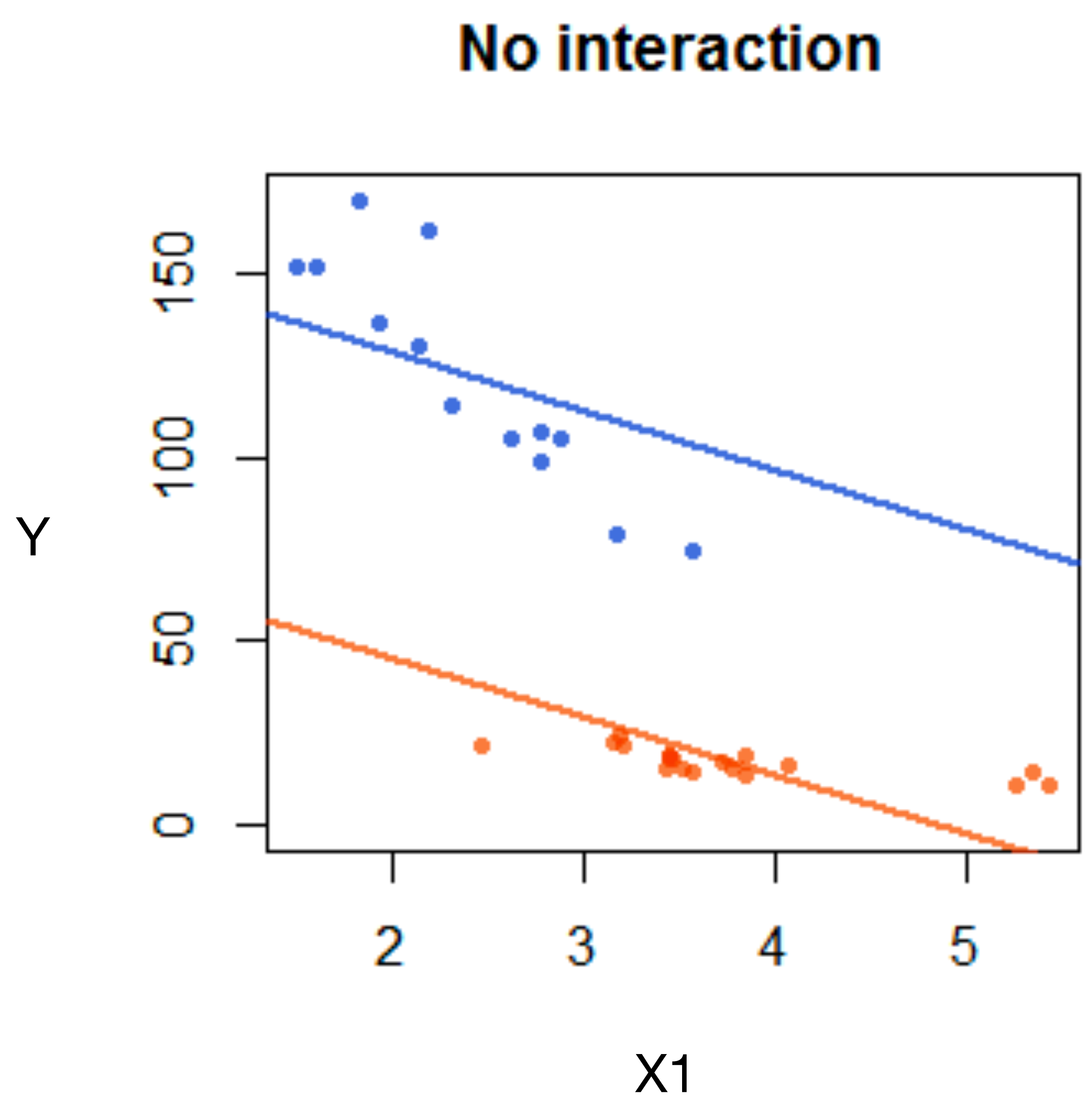
- Assume a model of the form: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_p X_p + \epsilon$
- Good for answering:
  - Is at least one of the X_j useful in predicting Y?
  - Are all p predictors necessary, or will only a subset suffice?
  - How accurate are the predictions? How well does the model fit?



Multiple linear regression

- What if there is a relationship between the predictors? What is a non-linear relationship is present?

- Can add **interaction** terms:



$X_2 = 0$ if male
$X_2 = 1$ if female

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

M: $Y = \beta_0 + \beta_1 X_1 + \epsilon$

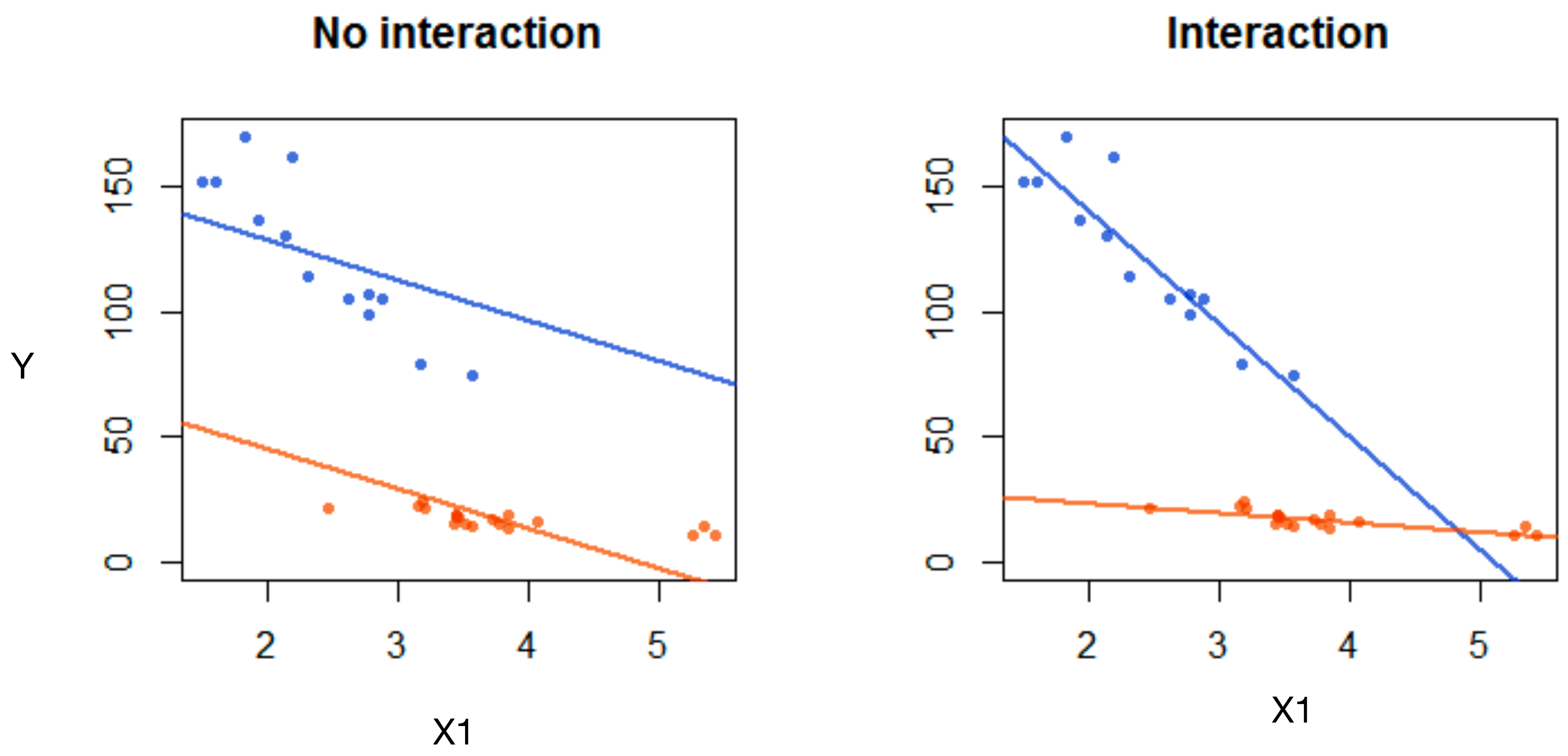
F: $Y = \beta_0 + \beta_1 X_1 + \beta_2 + \epsilon$

Multiple linear regression

- What if there is a relationship between the predictors? What is a non-linear relationship is present?

- Can add **interaction** terms:

$X_2 = 0$ if male
$X_2 = 1$ if female



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

M: $Y = \beta_0 + \beta_1 X_1 + \epsilon$

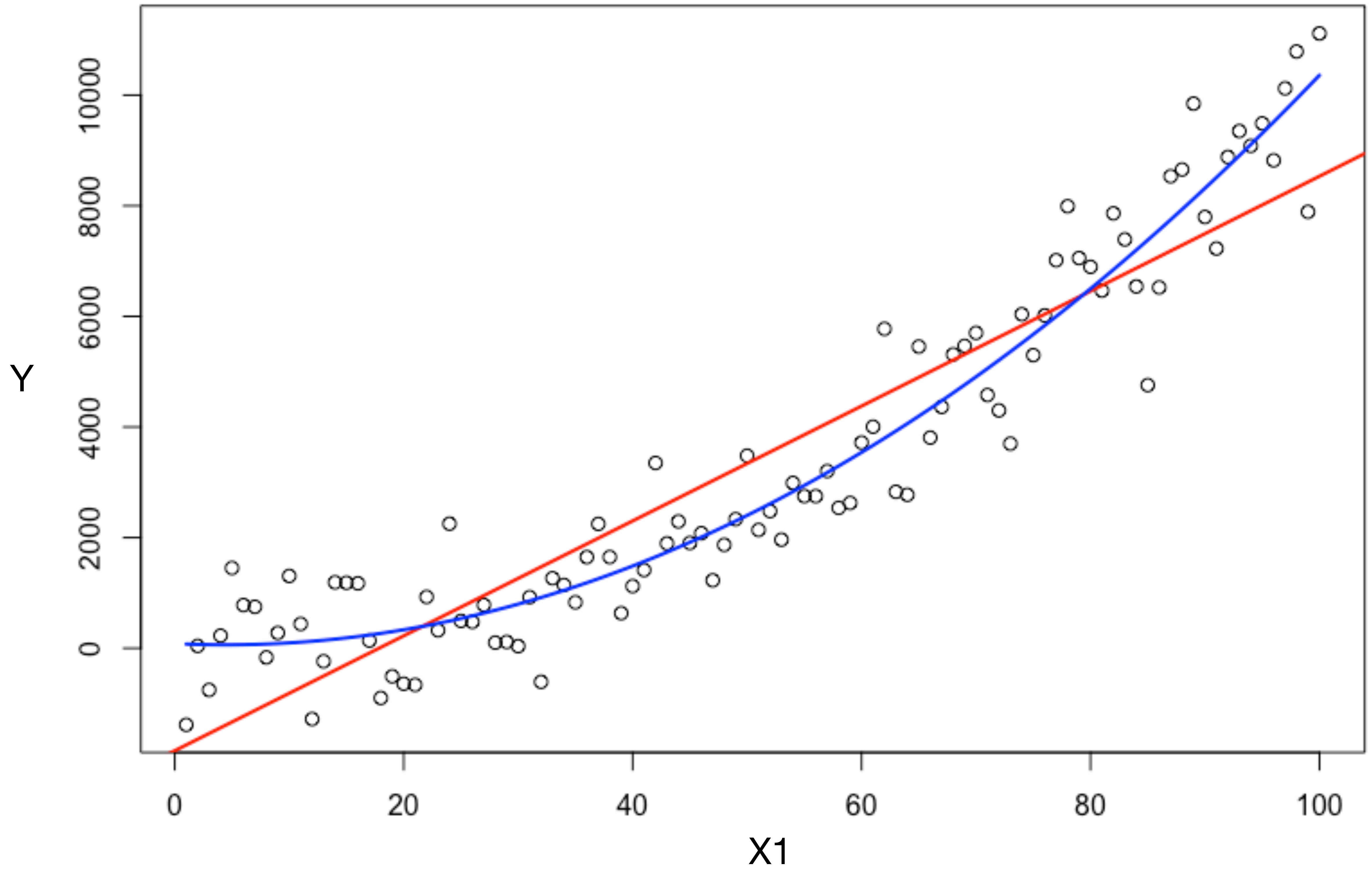
F: $Y = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3 X_1 + \epsilon$

Multiple linear regression

- What if there is a relationship between the predictors? What is a non-linear relationship is present?
- Can add **interaction** terms or **higher order** terms:

Red curve: $Y = \beta_0 + \beta_1 X_1 + \epsilon$

Blue curve: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$



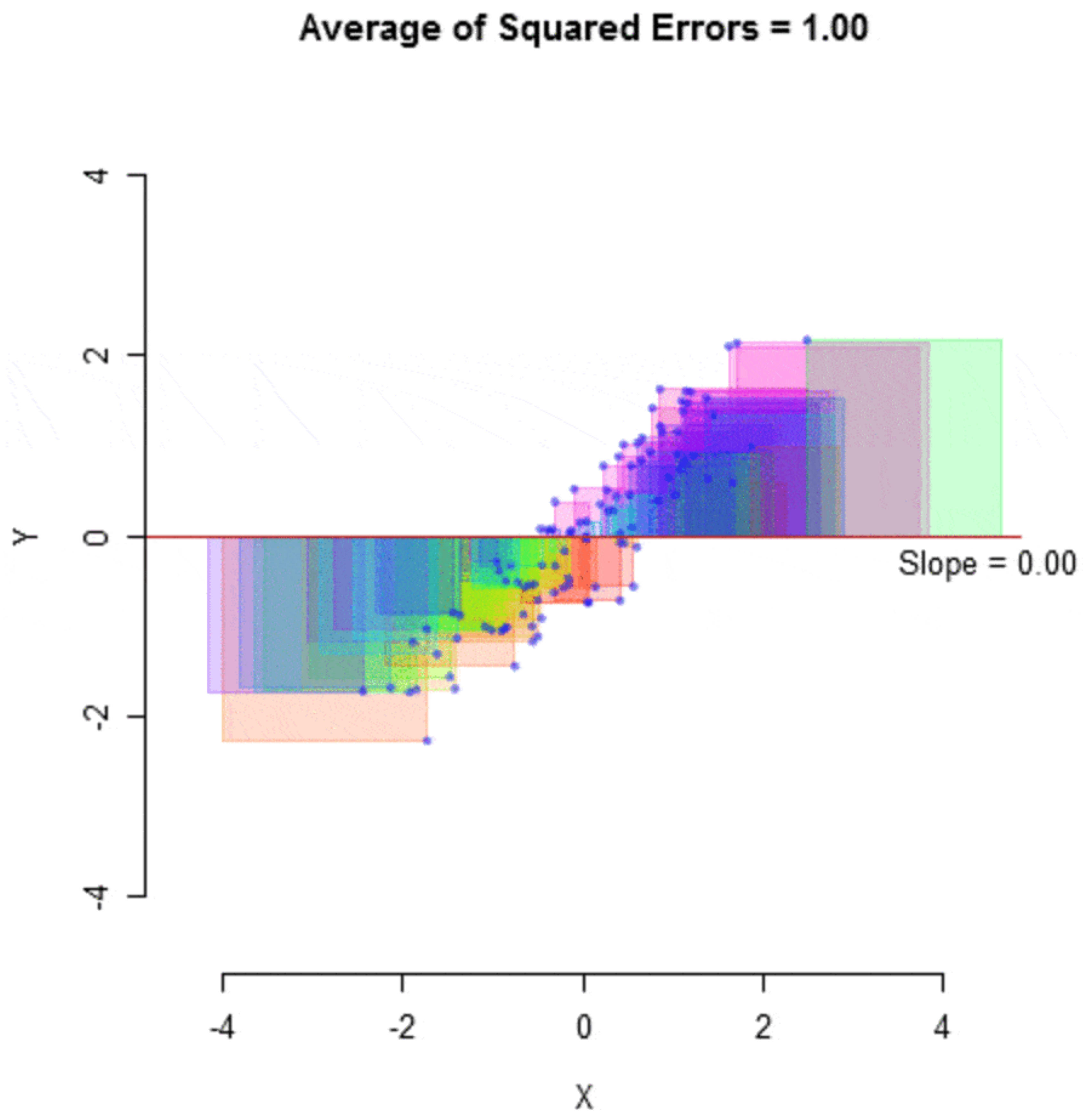
Is the blue model still **linear** regression?

Least squares for fitting regression models

- MLR: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_p X_p + \epsilon$
- Choose the $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that minimize the sum of squared residuals:

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

$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip})^2$$



What happens when you have many predictor variables?

- Two problems with high dimensional data

    1. **Interpretability**: Hard to summarize conclusions from a model with 100,000 predictor variables. Even harder when you include first, second, third, etc. order interactions

	Estimate
(Intercept)	0.2156926
dataX1	-0.2141724
dataX2	-0.5122957
dataX3	0.1171434
dataX4	0.0719710
dataX5	0.8538264
dataX6	0.7151485
dataX7	0.2723380
dataX8	-0.6684414
dataX9	0.7645613
dataX10	-0.8755725
dataX11	-0.1164399
dataX12	0.7995498
dataX13	0.0389759
dataX14	-0.8872863
dataX15	-0.3894061
dataX16	0.1546200
dataX17	-0.2036463
dataX18	0.0699375
dataX19	0.3227768
dataX20	0.0712609
dataX21	0.1855488
dataX22	0.1165093
dataX23	-0.6650188
dataX24	0.6449516
dataX25	-0.1417421

dataX73	-0.2317563
dataX74	0.0108767
dataX75	1.2230716
dataX76	-0.1464685
dataX77	0.2441201
dataX78	0.6476761
dataX79	1.1774111
dataX80	-0.7780623
dataX81	0.5206766
dataX82	0.7919490
dataX83	0.3892354
dataX84	-0.5359459
dataX85	0.6392728
dataX86	-0.6506848
dataX87	0.5911019
dataX88	-0.0154343
dataX89	1.0198047
dataX90	-1.0254036
dataX91	0.6058202
dataX92	-0.7472141
dataX93	0.0364057
dataX94	-0.0780022
dataX95	-0.0302979
dataX96	-0.3069039
dataX97	1.1033568
dataX98	-0.3277939
dataX99	-0.2405781

dataX206	-0.8356072
dataX207	-0.0950663
dataX208	-0.6155767
dataX209	0.1644539
dataX210	0.3324333
dataX211	0.0987915
dataX212	-0.6526612
dataX213	0.4841048
dataX214	0.3967542
dataX215	0.5935250
dataX216	-1.0240238
dataX217	0.1890421
dataX218	1.0827865
dataX219	0.1128421
dataX220	0.2807738
dataX221	-0.8270341
dataX222	-1.7440725
dataX223	-0.5586615
dataX224	0.0805911
dataX225	-0.3311416
dataX226	0.2456106
dataX227	-0.8335148
dataX228	-0.0895120
dataX229	-0.2507370
dataX230	-0.0087415
dataX231	-0.3336044
dataX232	0.4398568

What happens when you have many predictor variables?

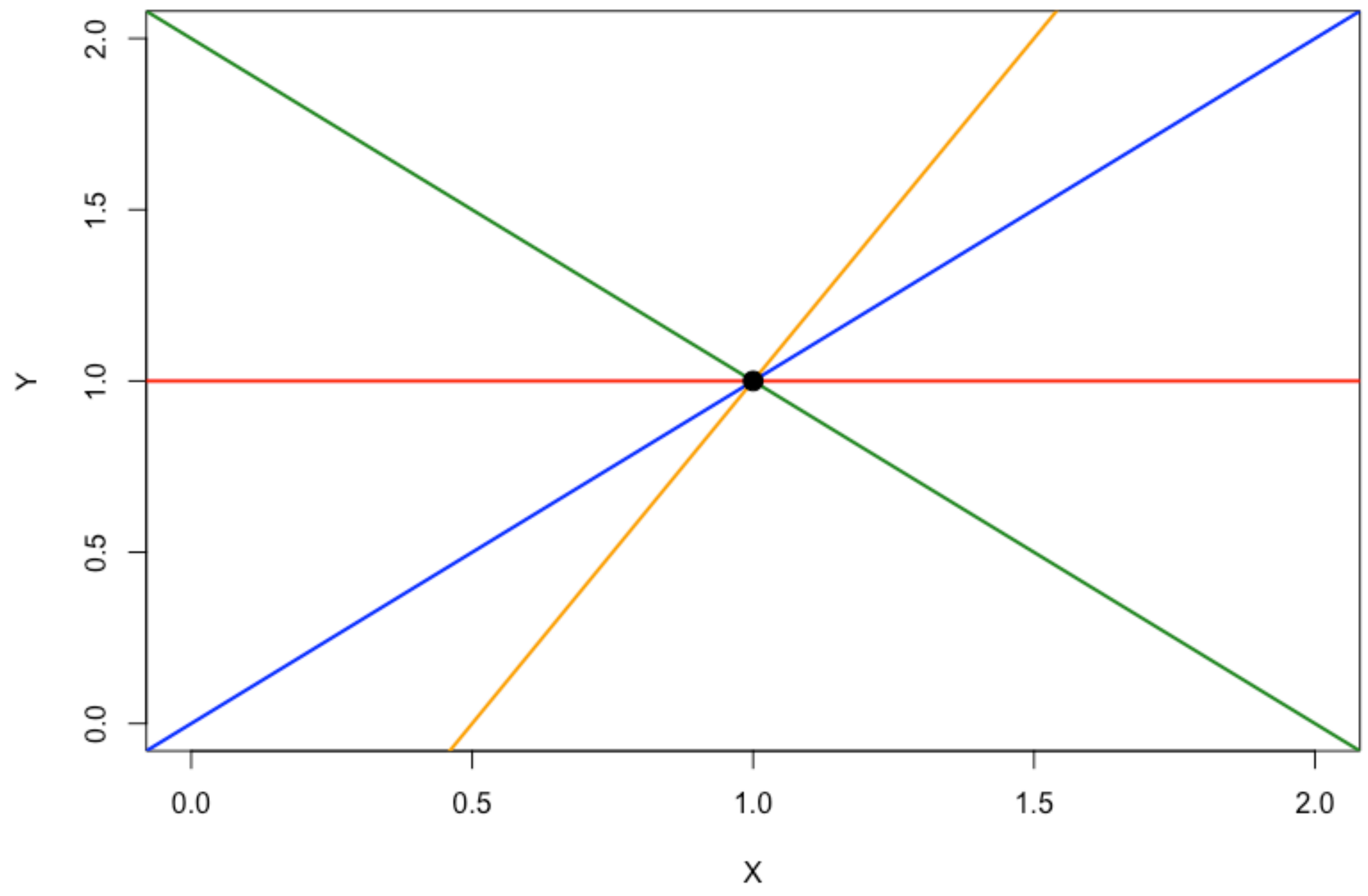
- Two problems with high dimensional data

  1. **Interpretability**

  2. **Prediction accuracy**:

     - If number of observations (n) is not much larger than number of predictor variables (p), then least squares fit can have high variability.

     - If n < p, then least squares fit does not have unique solution (infinite variance!)



Regularization for model fitting

- Regularization helps with:
    - **Interpretability**: certain methods can perform variable selection (setting coefficient estimates to EXACTLY zero)
    - **Prediction accuracy:** shrinks estimated coefficients towards zero (this reduces model variability)

Note: regularization has the same goal as least squares -> estimate $\beta_0, \ldots, \beta_p$

Regularization for model fitting

- Regularization helps with:
  - **Interpretability**: certain methods can perform variable selection (setting coefficient estimates to EXACTLY zero)
  - **Prediction accuracy:** shrinks estimated coefficients towards zero (this reduces model variability)

- Example: the LASSO coefficient estimates $\hat{\beta}_0, \ldots, \hat{\beta}_p$ minimize:

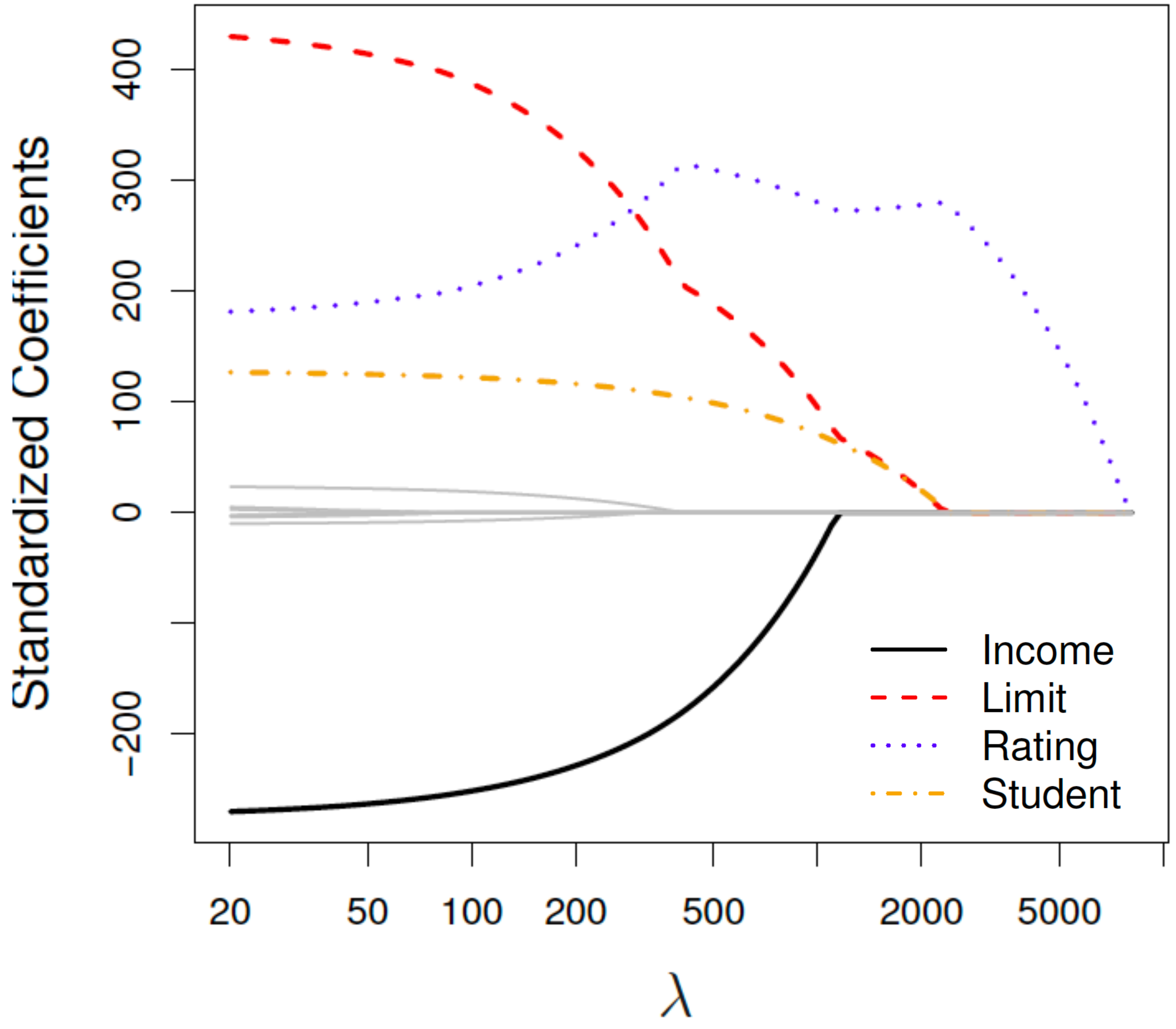
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

Controls how well model fits the data

Controls overall magnitude of coefficients

Balance between model fit and coefficient magnitude balanced by λ

What role does λ play?



- λ balances model fit and size of coefficient estimates

- How to pick λ?

    - Test many different λ options, pick the one that optimizes a performance measure

    - E.g., adjusted R^2, BIC, AIC, out of sample prediction error

Regression summary

- Linear regression
  - easy to implement
  - provides advantages in terms of interpretability and inference compared to non-linear methods
  - Methods for fitting parameters that we covered
    1. Least squares (minimize residual sum of squares)
    2. Regularization (minimize balance between RSS and size of coefficients)
       - Can improve least squares fit by reducing complexity
       - Use it when you have many predictor variables
       - It still assumes a linear model

Other regression methods

- Linear assumption can only go so far! Other methods for regression:
  - Smoothing splines
  - K-nearest neighbors
  - Tree-based methods

All of these address the same problem!

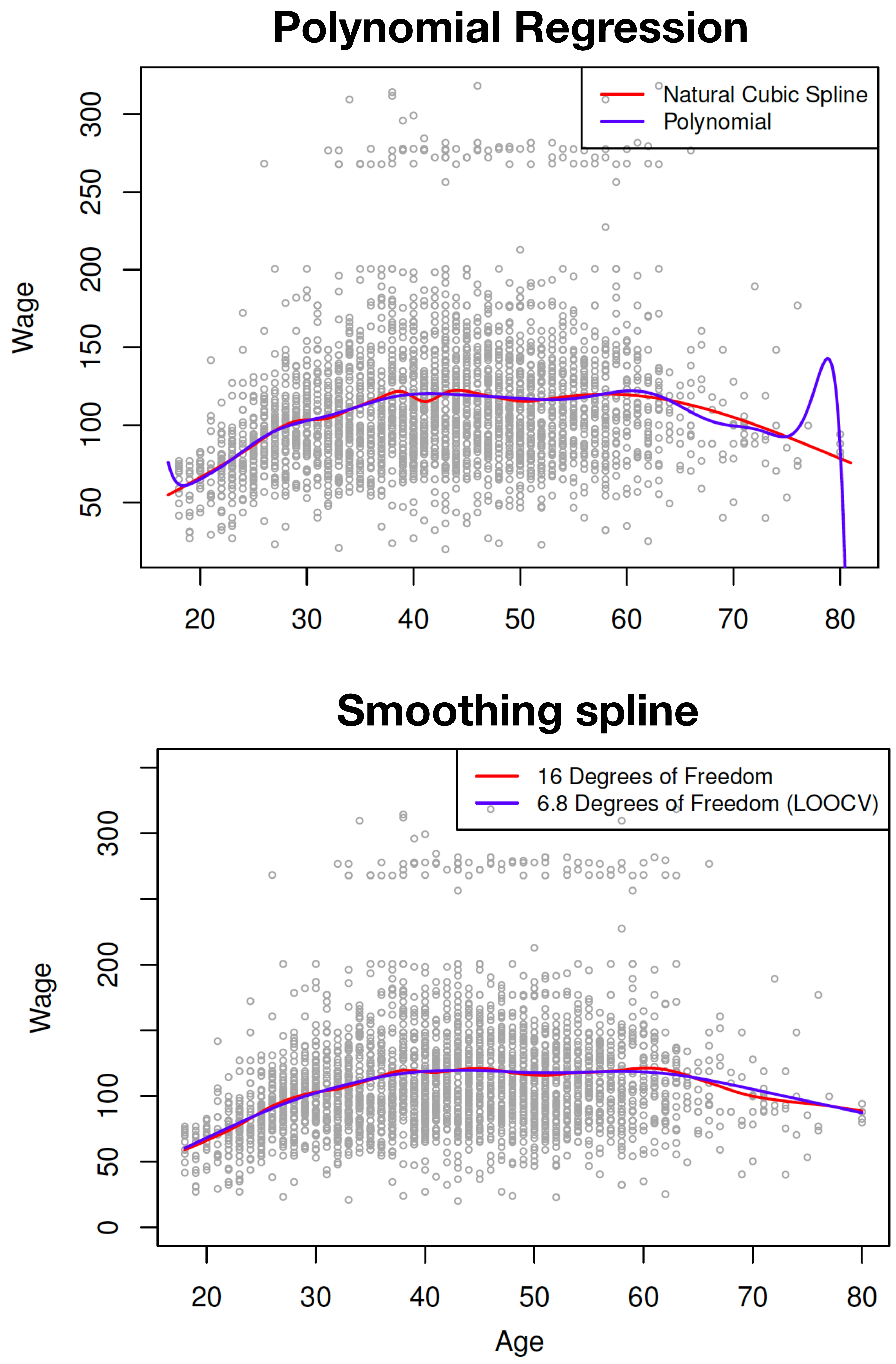
Try to estimate relationship between X and Y

Smoothing splines

- Want a smooth curve, $g(x_i)$, that fits the data well.
  - That is, minimize $\sum_{i=1}^{n} (y_i - g(x_i))^2$
  - Without constraint, $g(x_i)$ will interpolate!

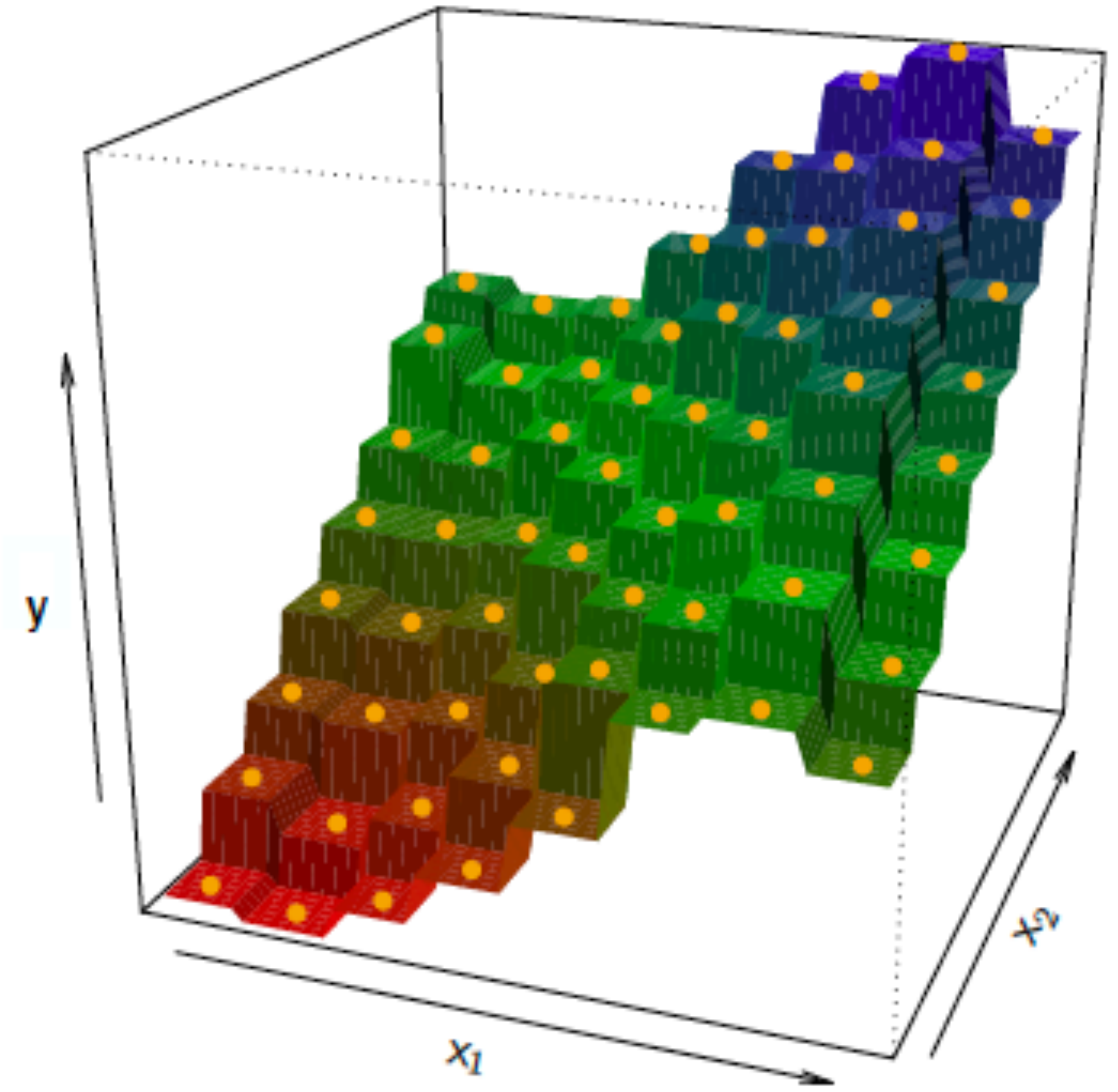
- Pick g that minimizes: $$\sum_{i=1}^{n} (y - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

- $\int g''(t)^2 dt$ is a measure of total change in the function $g'(t)$

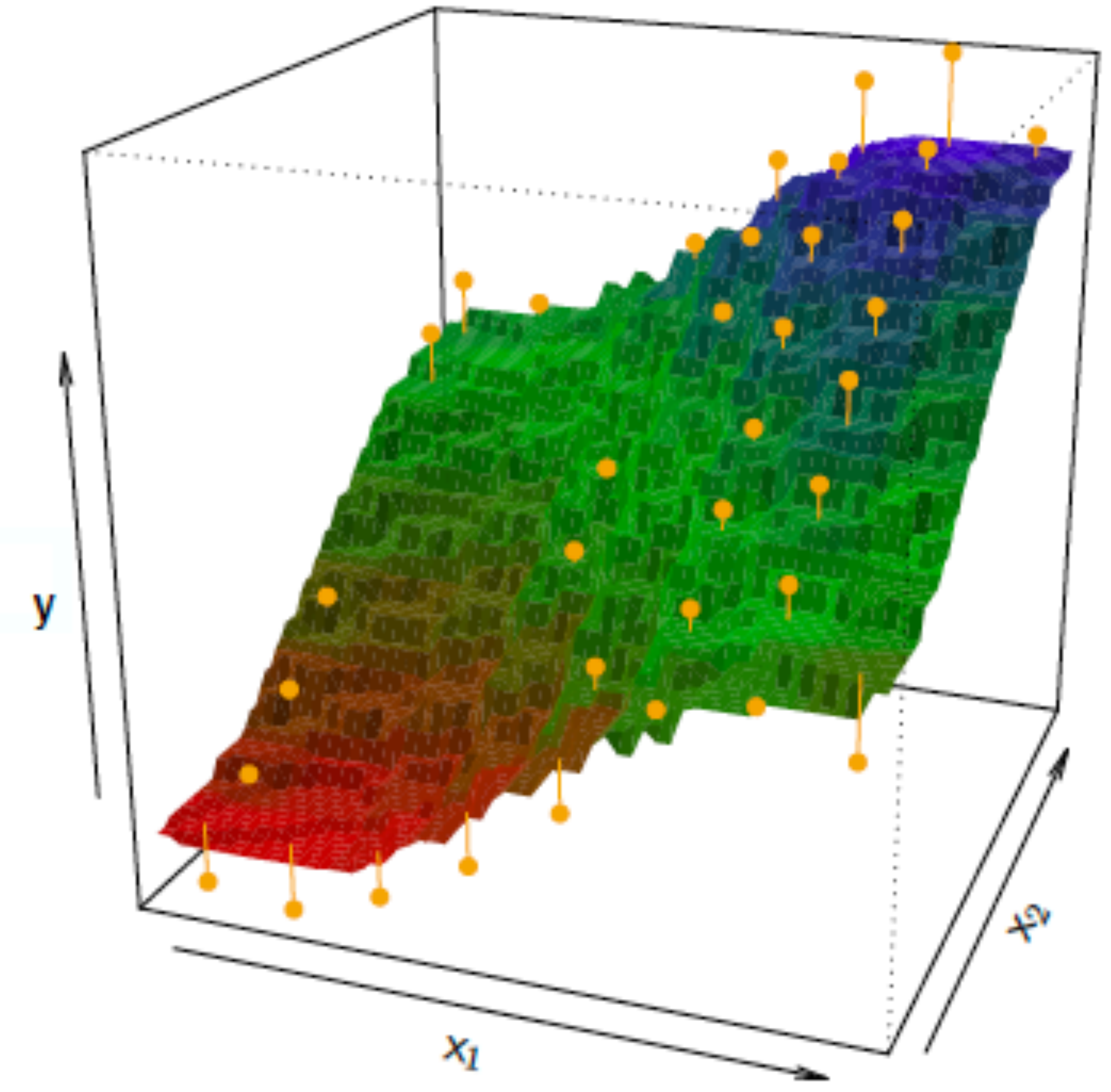


K-nearest neighbors

- Non-parametric method
- Parametric approach (linear regression) tends to outperform non-parametric approach (KNN) when selected model form is close to the true relationship
- You want a prediction of Y at some set of predictor variable values, X_0.
  - Pick a value K
  - KNN returns the average of the corresponding response values (Y) of the K closest data points to X_0
- Be careful with high dimensions!



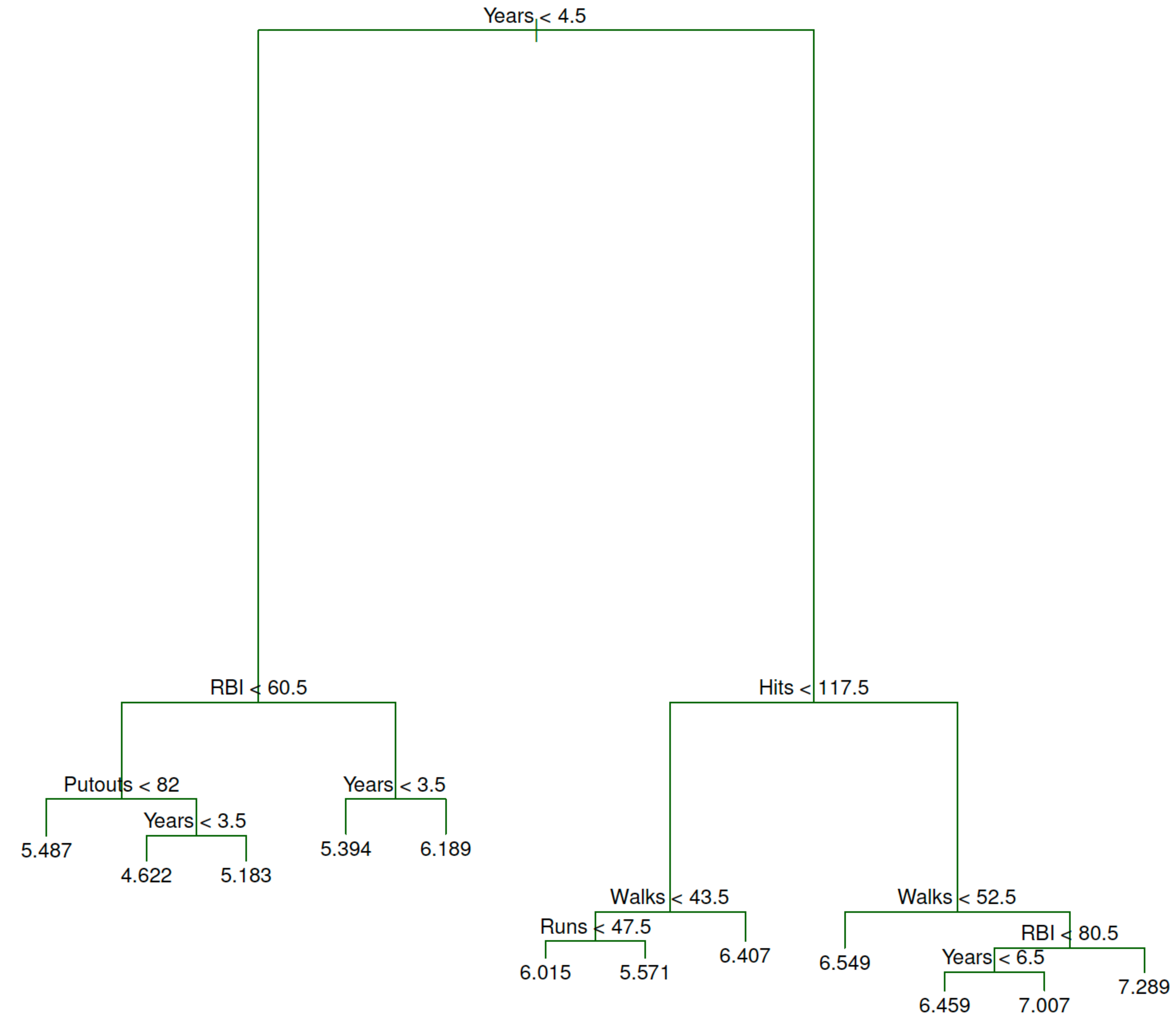
K=1

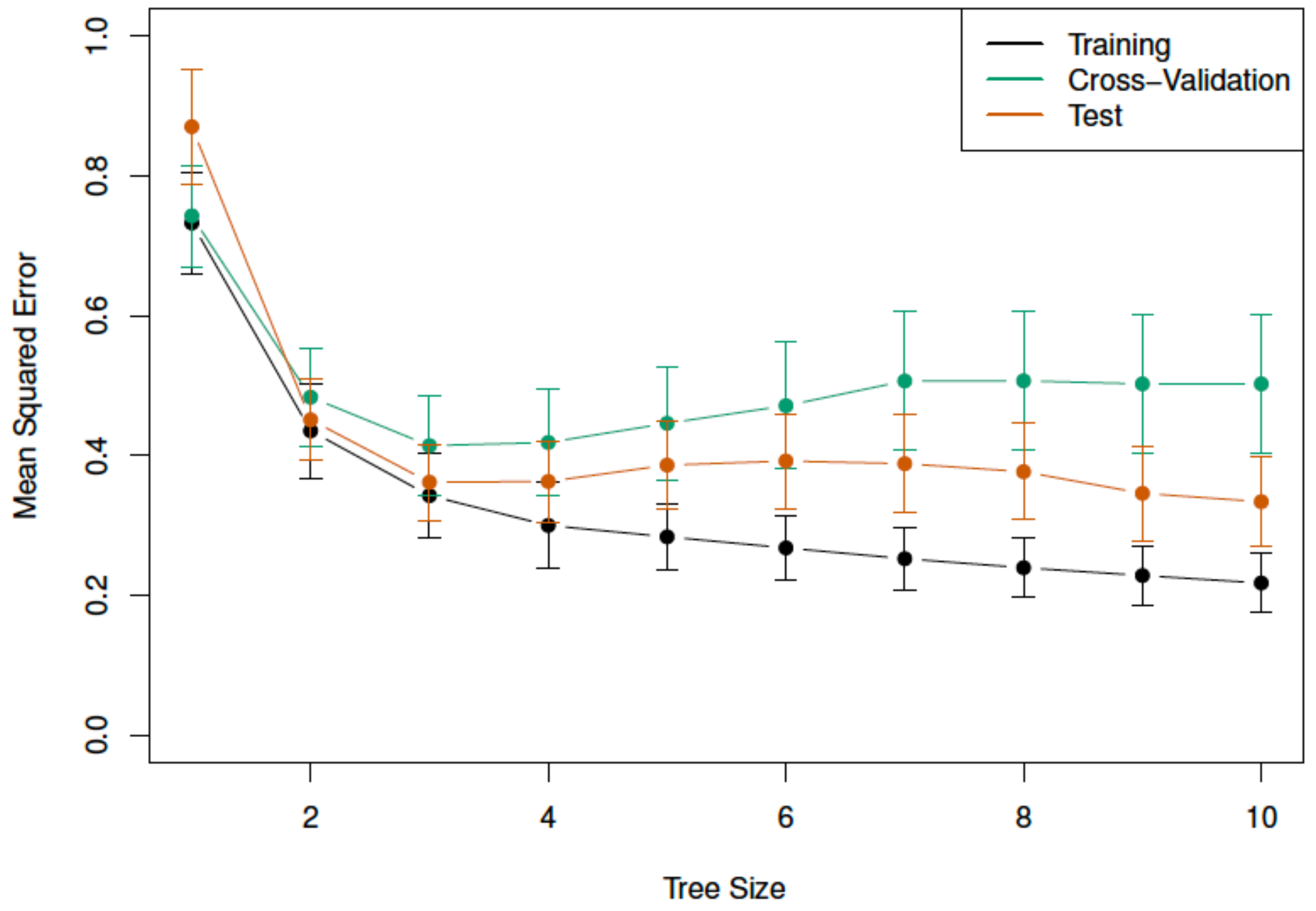


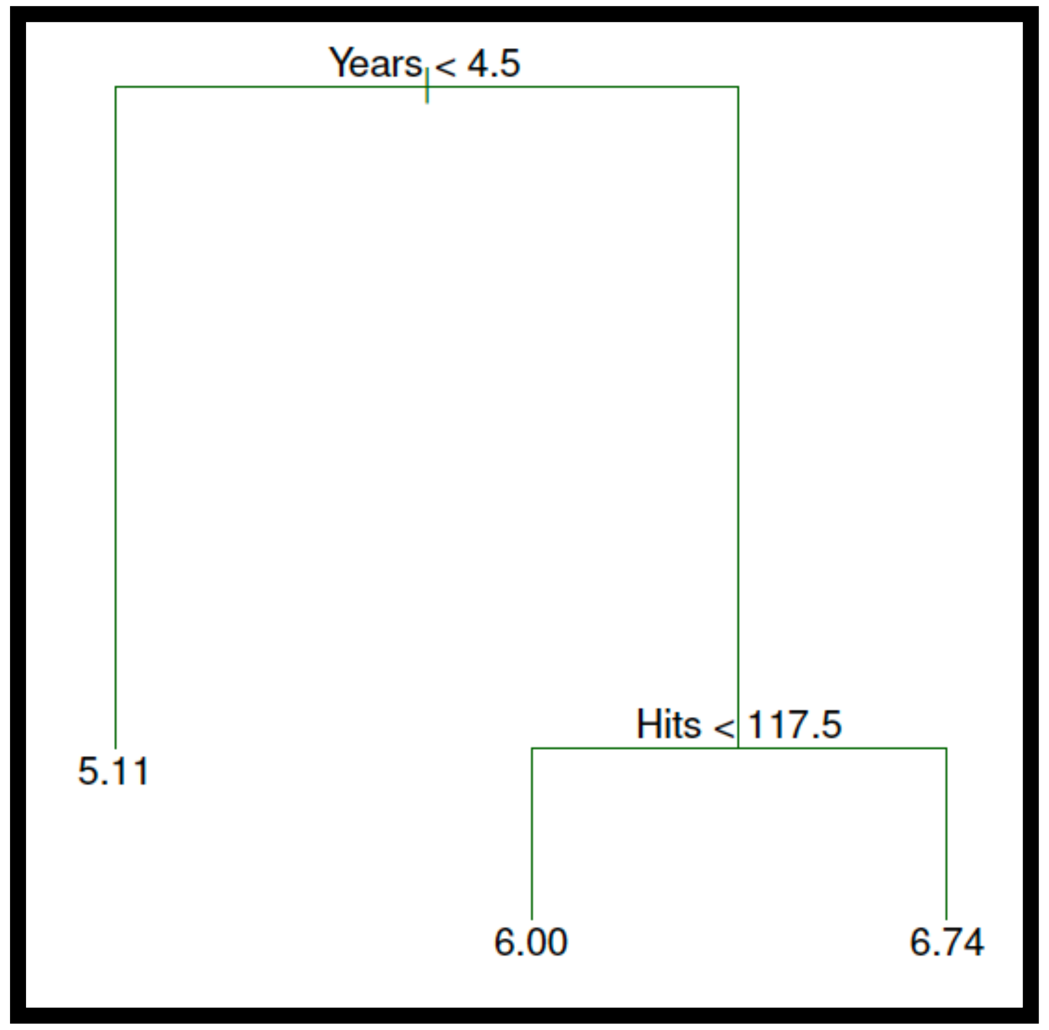
K=9

Tree-based methods

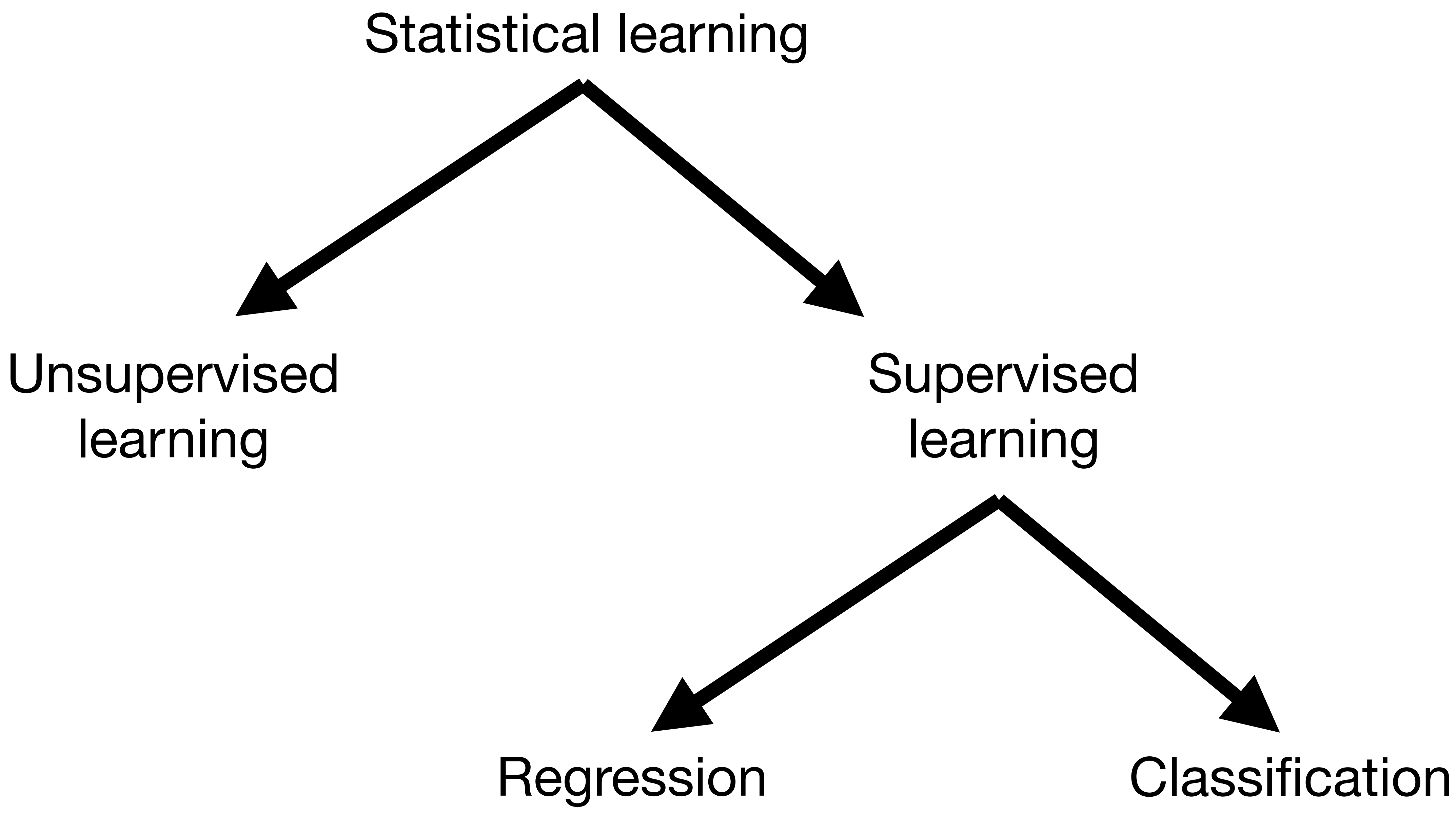
- Simple and easy to interpret
- Segment the predictor space into regions
- To make a prediction for x_0, use the mean response for the training observations in the region to which x_0 belongs
- "Top-down, greedy" method used to fit the full tree
- Use CV to go back and "prune" the full tree to reduce variability







Classification



Classification problem overview

- We observe measurements $X = (X_1, X_2, \ldots, X_p)$ and associated response Y that takes **categorical** values

- Why not encode the categories in Y as numbers and use linear regression?

$$Y = \begin{cases} 1 & \text{if } \texttt{stroke}; \\ 2 & \text{if } \texttt{drug overdose}; \\ 3 & \text{if } \texttt{epileptic seizure}. \end{cases}$$

- Will discuss three methods for classification: logistic regression, KNN, tree-based

Logistic regression

- Consider a categorical Y with two options: Yes or No

- Interested in modeling $p(X) = \Pr(Y = \text{Yes} \mid X)$, where $X = (X_1, \ldots, X_p)$

- Logistic regression similar to linear regression, but model output restricted to [0,1]

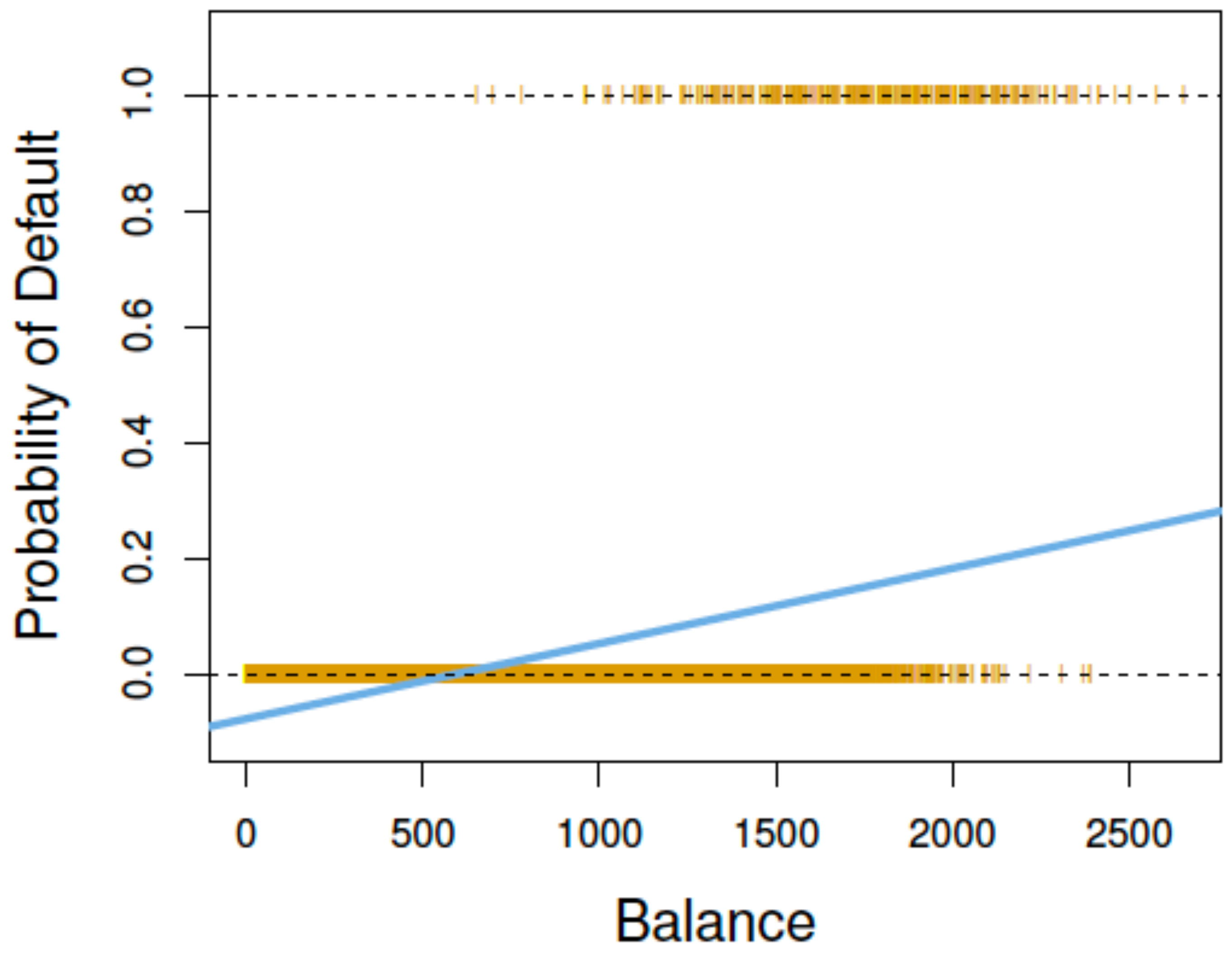
- Use the logistic function:

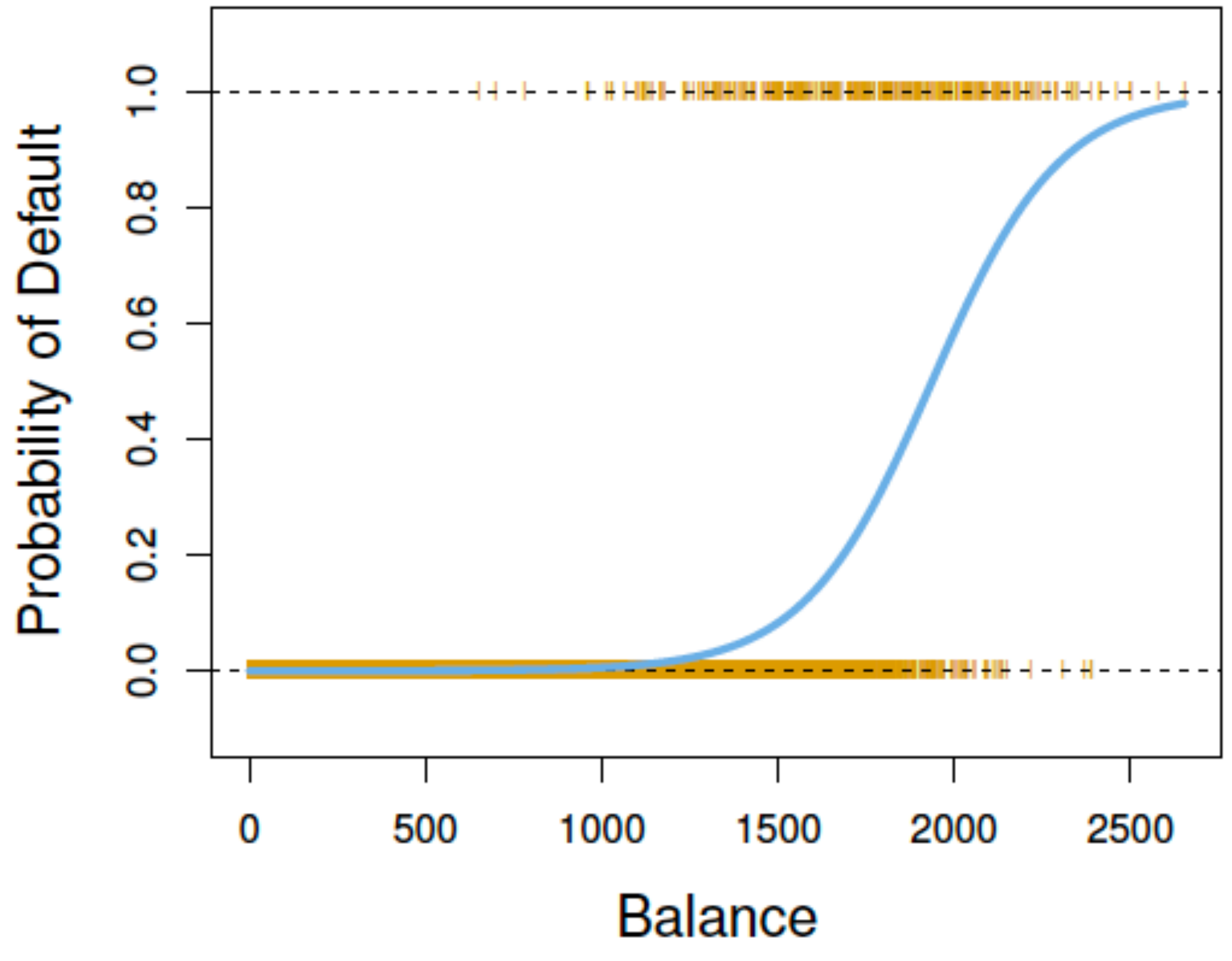
$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p}}$$

- Fit coefficients using “maximum likelihood”

- To make predictions, set some threshold on $p(X)$ to distinguish Yes/No

- Extensions available for categorical response that take > 2 values

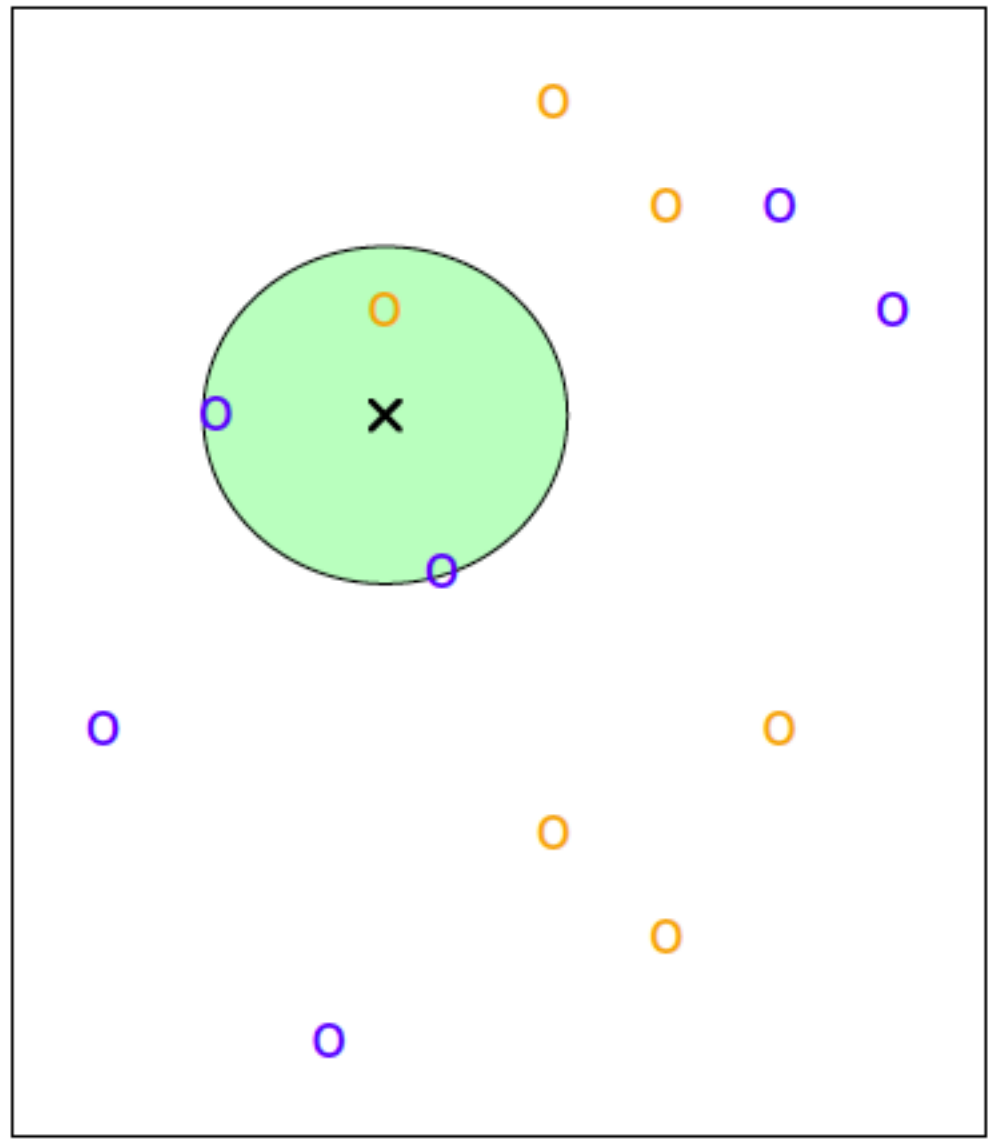


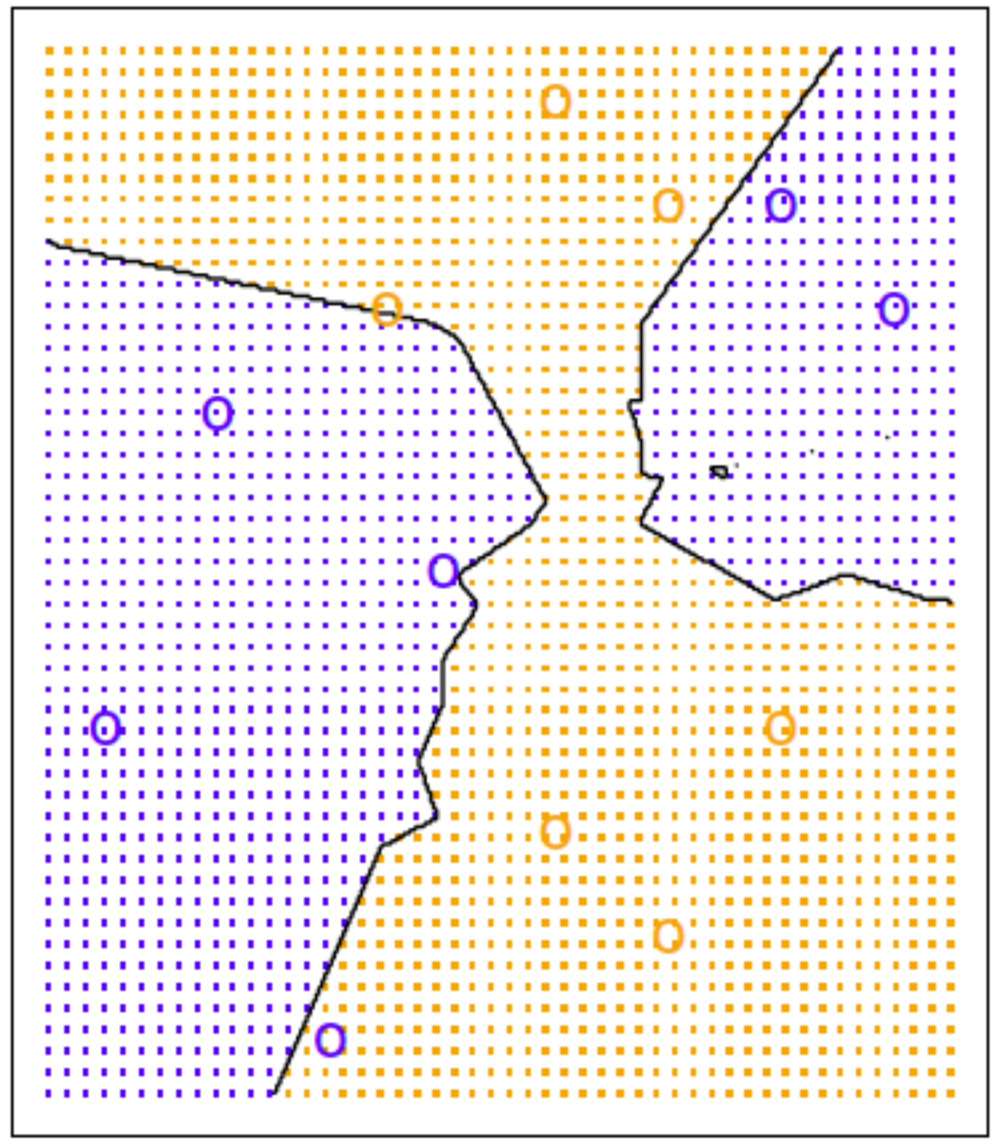


K-nearest neighbors

- Very similar to KNN in a regression setting
- Given a value K and prediction point x_0
  - KNN sets the class of x_0 to be the most common class in $\mathcal{N}_0$
  - where $\mathcal{N}_0$ are the K training observations closest to x_0

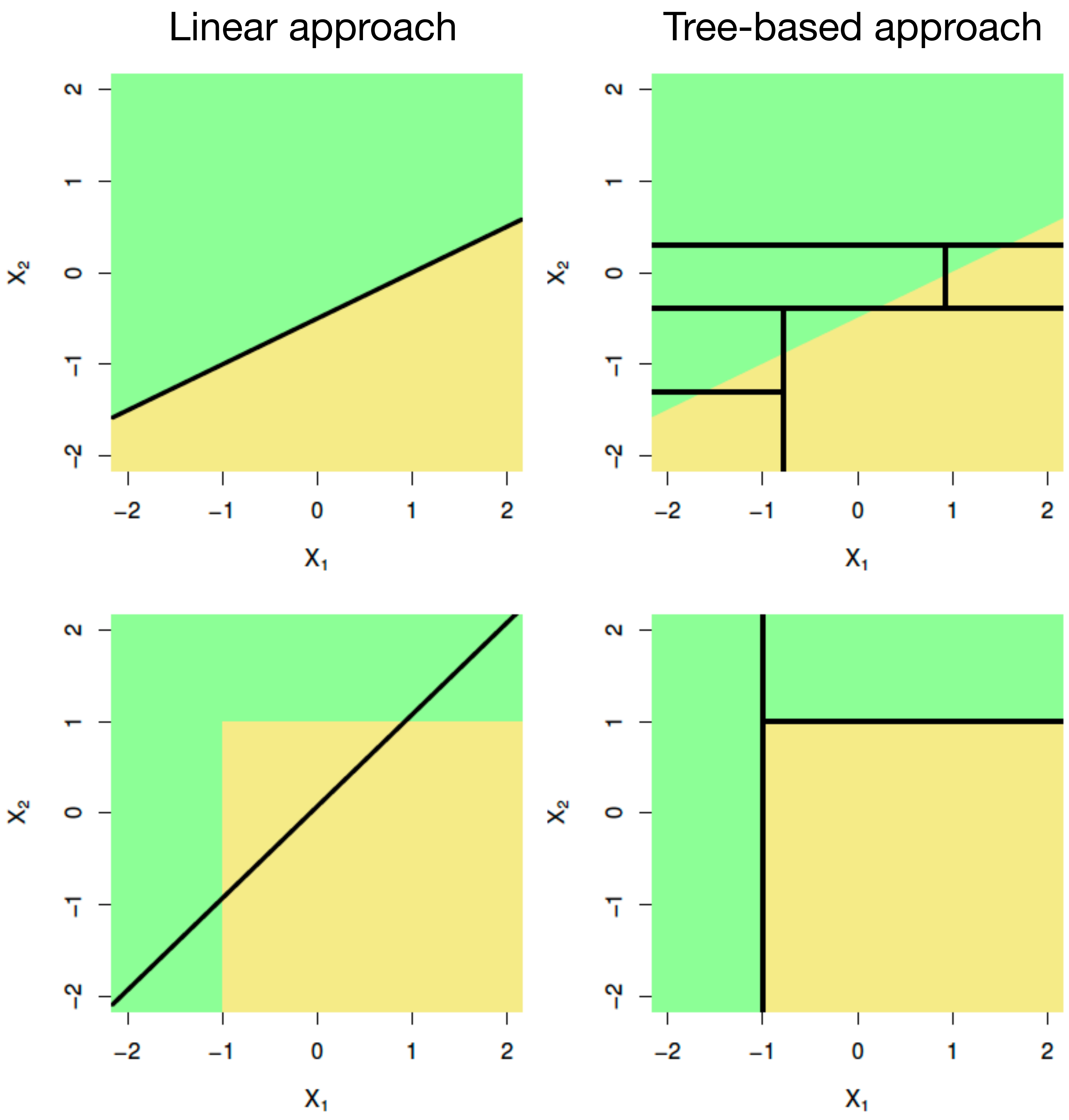
K=3





Tree-based methods

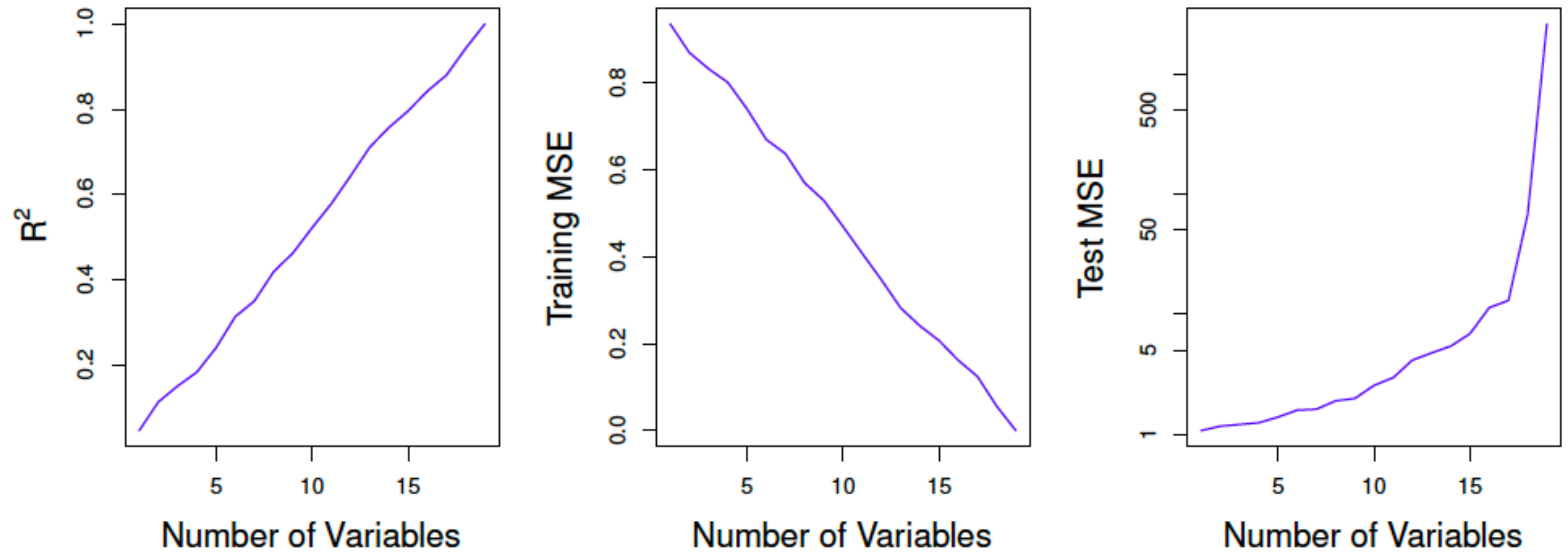
- Classification tree very similar to regression tree
- Segment the predictor space into regions
- To make a prediction for x_0, use the most common class for the training observations in the region to which x_0 belongs



Model Evaluation

General considerations

- Does it make sense to use a statistical model? How much data is available for training? Would a mechanistic model be better suited?
- Regression vs. classification?
- Does their model violate any assumptions? Are they using a model in a sub-optimal setting (e.g., KNN with large p)?
- Are the performance metrics suitable?

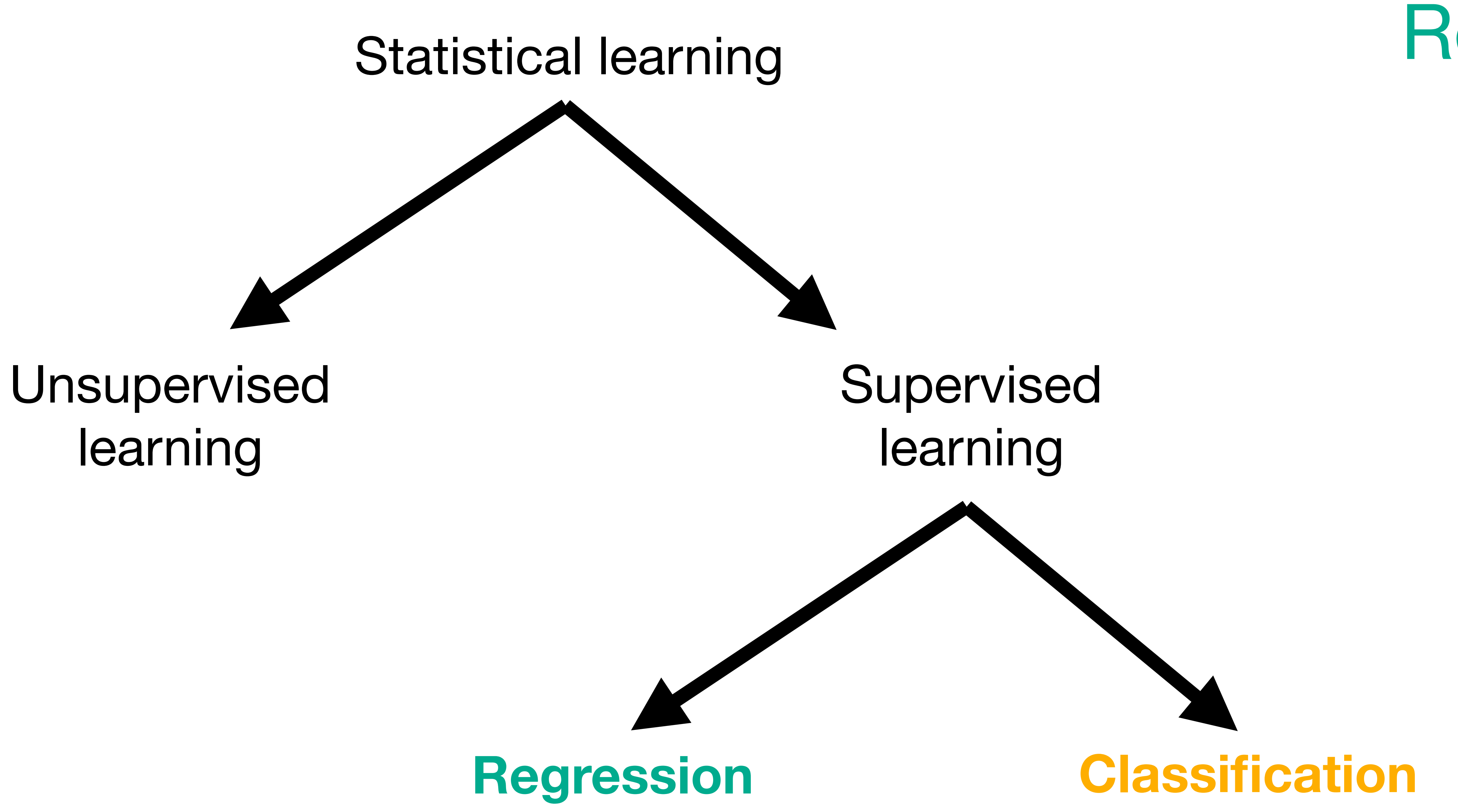


Simulated example with n = 20

R^2 and training MSE can make model look good when it is not!

Summary

Statistical learning involves building models to capture relationships in data



Regression

- Simple linear regression
- Multiple linear regression
- Smoothing splines
- K-nearest neighbors
- Tree-based methods

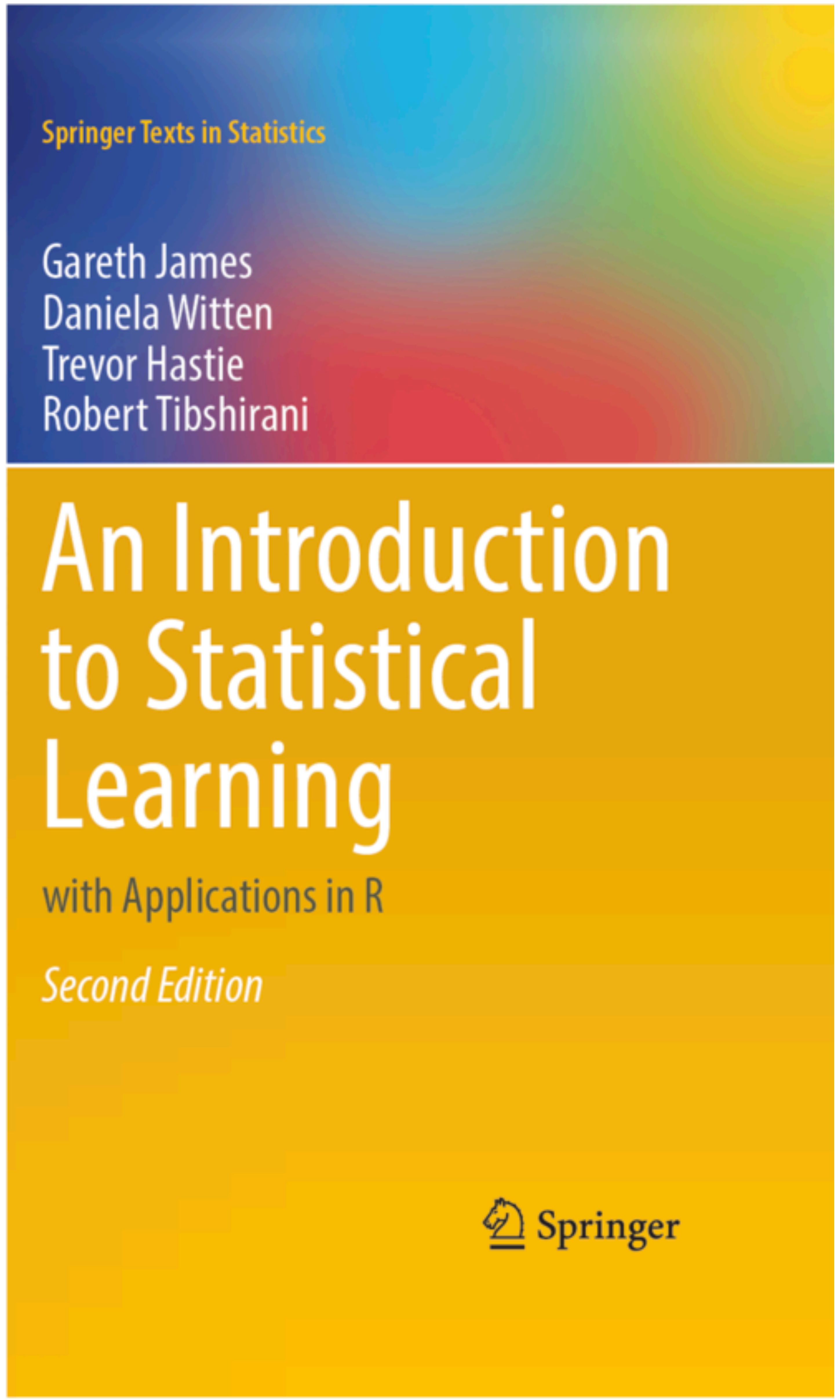
Classification

- Logistic regression
- K-nearest neighbors
- Tree-based methods

Which method to use? Check assumptions, then start simple and get more complex if necessary.

Great reference text

- Free pdf at: https://www.statlearning.com/
- Most of the images in this talk taken from ISLR

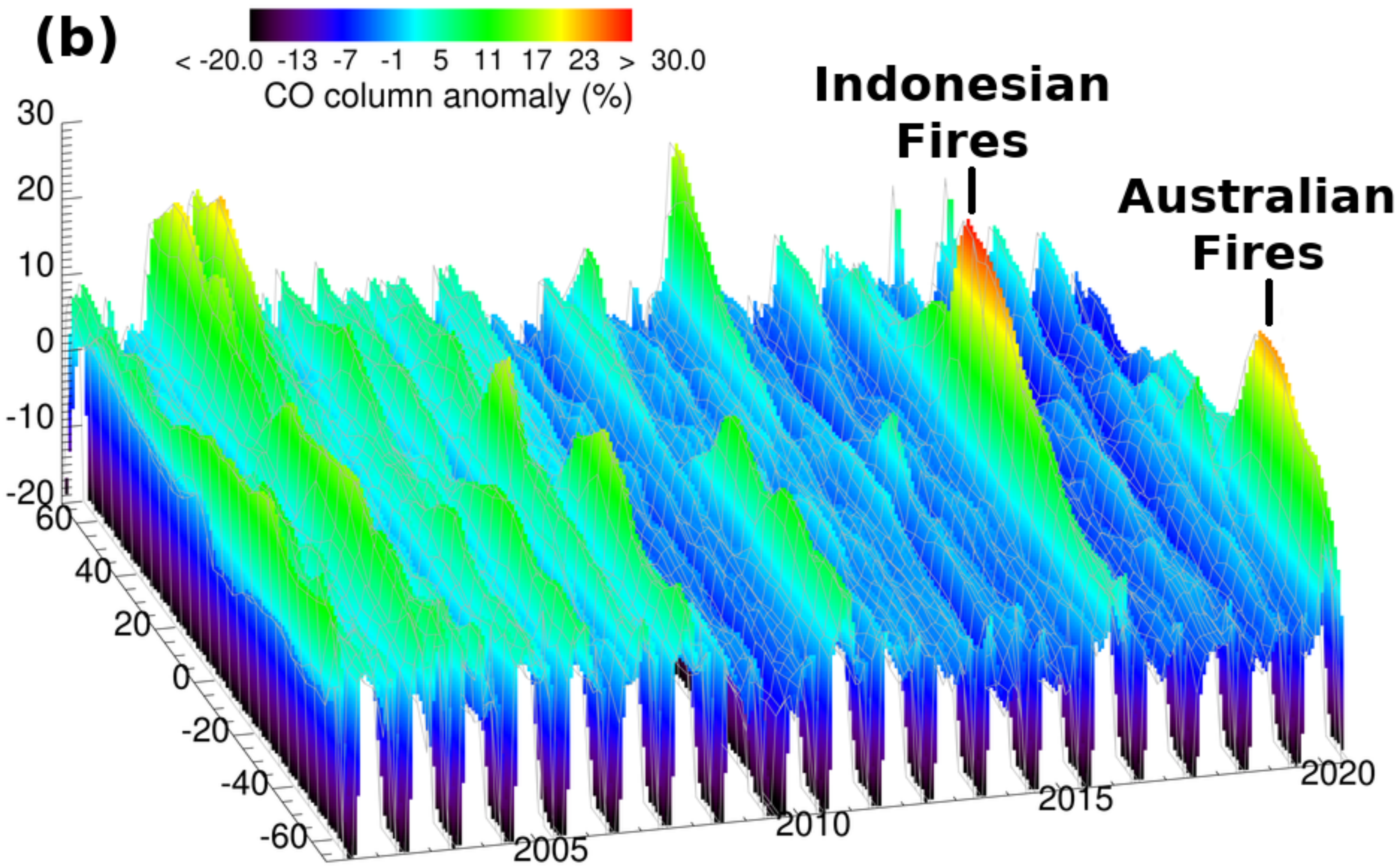


Statistical learning example:

What are the drivers of fire season intensity in Maritime Southeast Asia?

Motivation

Certain Southern Hemisphere regions experience extreme carbon monoxide (CO) anomalies as a result of biomass burning.



October 2015

Palangkaraya, Indonesia

January 2020

Canberra, Australia

Motivation

Certain Southern Hemisphere regions experience extreme carbon monoxide (CO) anomalies as a result of biomass burning.

October 2015

Palangkaraya, Indonesia

January 2020

Canberra, Australia

Our goals:

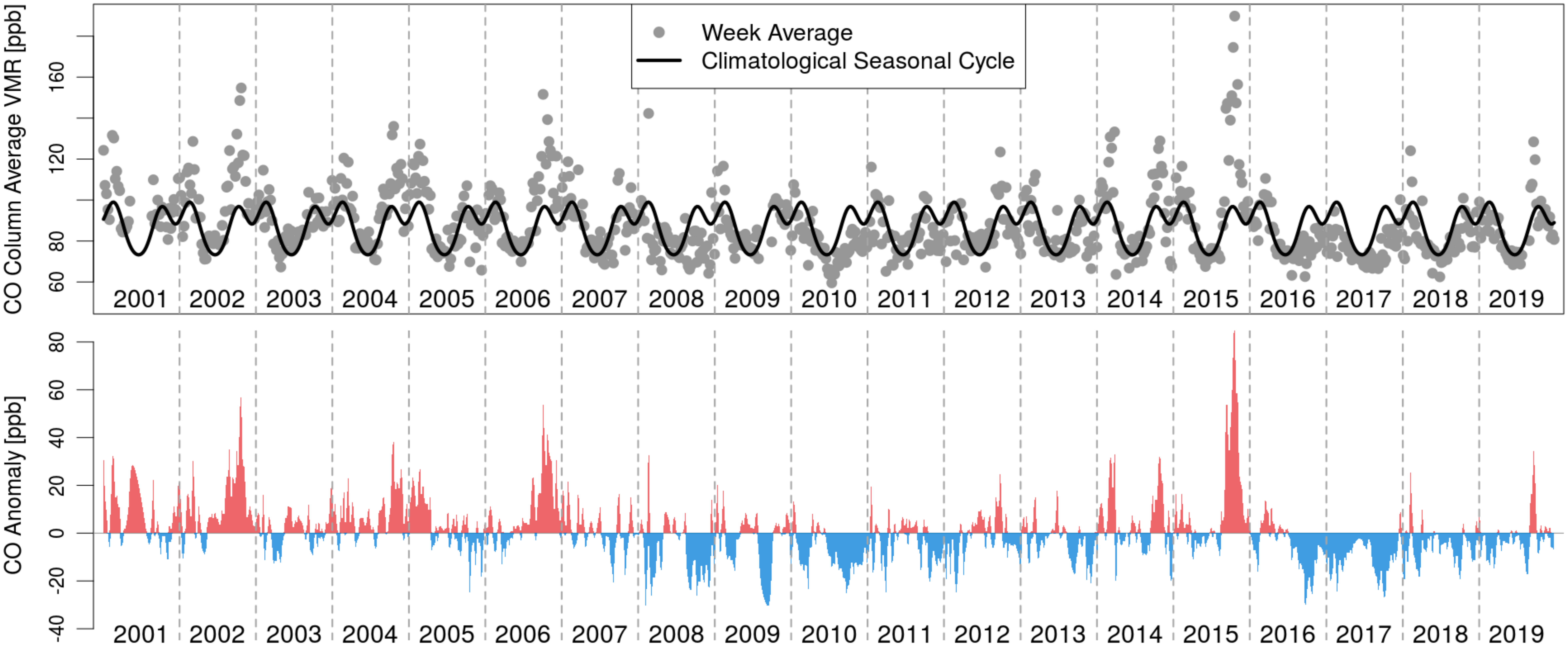
1. Predict CO at useful lead times
2. Build interpretable models for scientific conclusions

Response variable: carbon monoxide

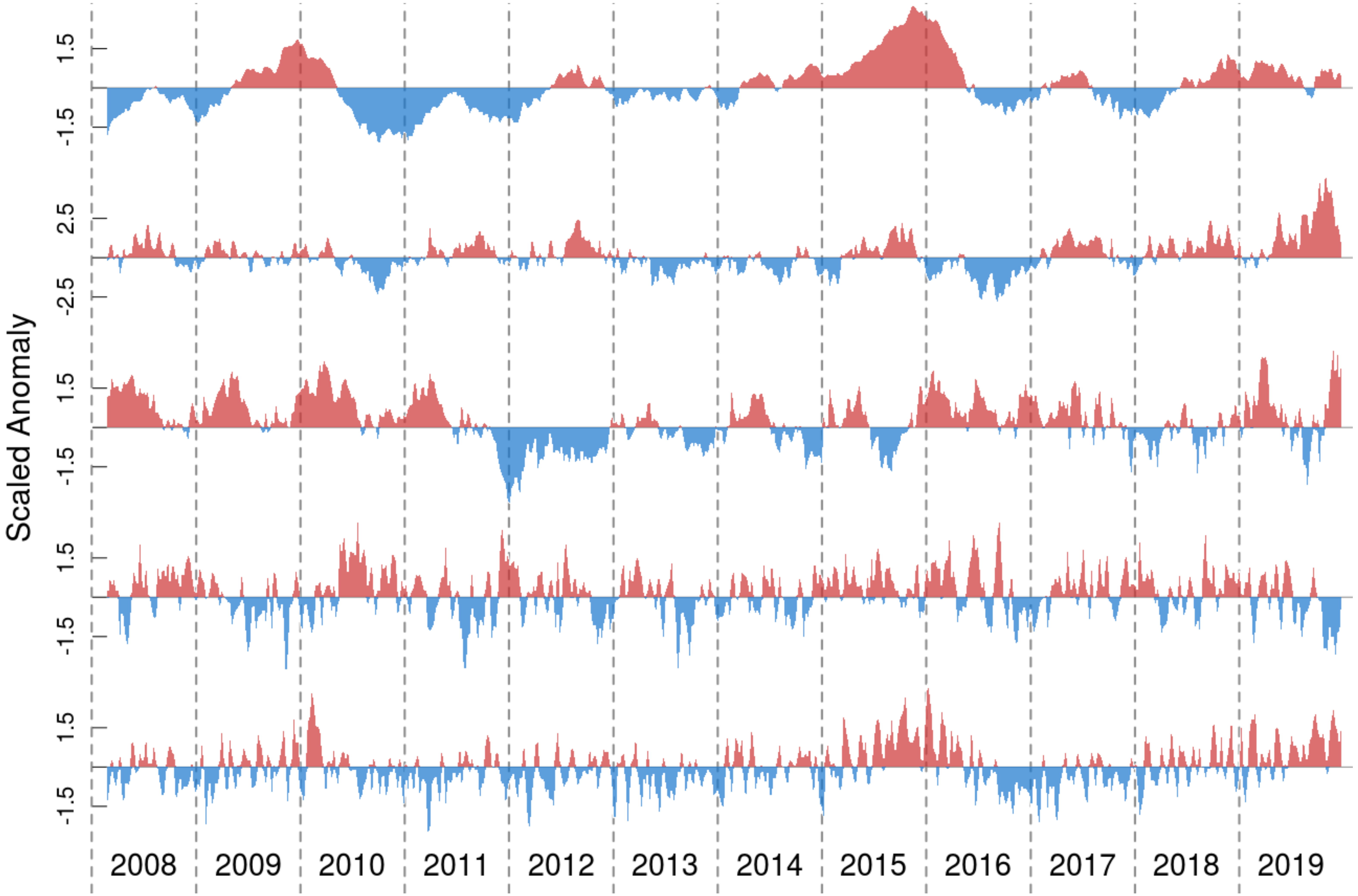
- Use multiple linear regression to model atmospheric CO
- CO aggregated within the MSEA biomass burning region via spatial and temporal averages

Response variable: carbon monoxide

Response variable: Deseasonalized, week-averaged CO anomalies at time t



Predictor variables: climate mode indices



Nino 3.4 (NINO)

Dipole Mode Index (DMI)

Tropical South Atlantic (TSA)

Antarctic Oscillation (SAM)

Outgoing Longwave Radiation (OLR)

Predictor variables: climate mode indices

Predictor variables: Week-averaged climate mode indices lagged at time t - τ

Statistical model

We use lagged multiple linear regression model with first order interactions and squared terms

$$CO(t) = \mu + \sum_{k} a_k \, \chi_k(t - \tau_k) + \sum_{i,j} b_{ij} \, \chi_i(t - \tau_i) \, \chi_j(t - \tau_j) + \sum_{l} c_l \, \chi_l(t - \tau_l)^2 + \epsilon(t)$$

Main effects — Interaction terms — Squared terms

CO(t) - CO anomaly in a given response region at time t

μ - constant mean displacement

χ - climate indices

τ - lag value for each index in weeks

ϵ(t) - error term

Regularization for variable and lag selection

We consider lags between 1 and 52 weeks for each index

- Results in far more covariates than observations
- Regularization well suited for this regime (p >> n)

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \sum_{j=1}^{p} p(\beta_j)$$

Regularization for variable and lag selection

We consider lags between 1 and 52 weeks for each index

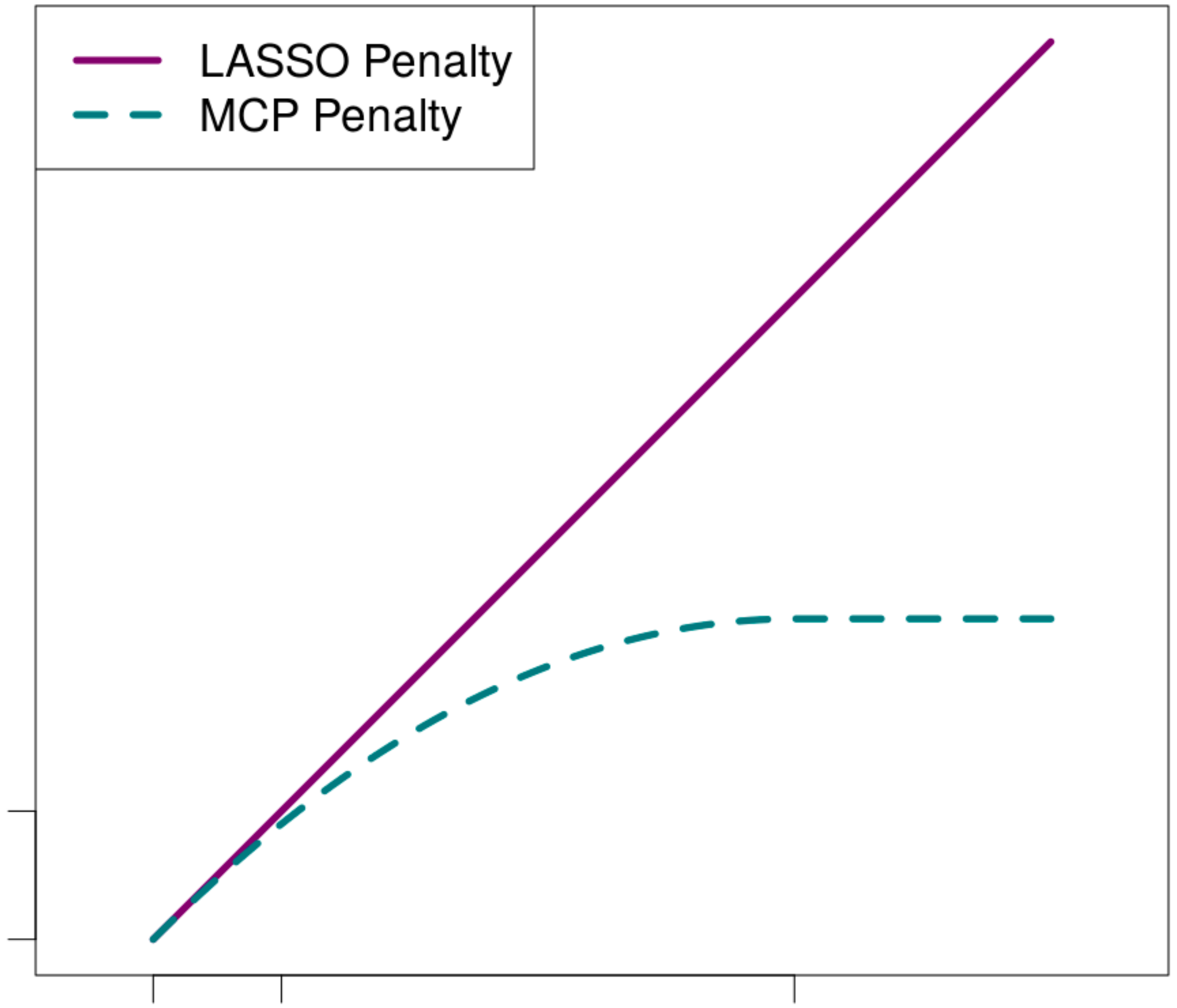
- Results in far more covariates than observations
- Regularization well suited for this regime (p >> n)

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \sum_{j=1}^{p} p(\beta_j)$$

We use the minimax concave penalty (MCP)

LASSO $p(\beta) = \lambda|\beta|$

MCP $p(\beta) = \begin{cases} \lambda|\beta| - \frac{\beta^2}{2\eta} & \text{if } |\beta| \leq \eta\lambda \\ \frac{\eta\lambda^2}{2} & \text{otherwise.} \end{cases}$



Regularization for variable and lag selection

Evaluate models along the solution path via the extended Bayesian information criterion (EBIC)

- Similar to BIC, but can increase penalty on larger models
- Control with free parameter $\gamma \in [0,1]$
- $\gamma \to 1$ results in smaller models
- $\gamma \to 0$ results in the BIC (and hence larger models)

Picking parameter values

- For a given γ, vary η and λ in a grid search
- Pick the model that minimizes EBIC for that γ
- More on γ selection to come!

Free parameters:

Regularization $\to \lambda$

MCP $\to \eta$

EBIC $\to \gamma$

Interpretable models lead to scientific conclusions

$$\gamma = 1$$

```
           Est (Std. Error)
(Intercept)    -1.6 (0.78)
nino_4          7.2 (0.78)
dmi_4           7.2 (0.93)
dmi_12         -8.0 (0.87)
aao_51         -3.1 (0.67)
olr_1           3.5 (0.79)
I(nino_4^2)     2.5 (0.54)
nino_4:olr_1    3.5 (0.76)
nino_4:dmi_12 -6.5 (0.77)
aao_51:olr_1   -2.3 (0.67)


Adjusted R-squared:   0.60
```

Smallest model highlights important climate-chemistry connections:

1. NINO has strong influence on CO at a four week lead time

Interpretable models lead to scientific conclusions

$$\gamma = 1$$

```
            Est (Std. Error)
(Intercept)    -1.6 (0.78)
nino_4          7.2 (0.78)
dmi_4           7.2 (0.93)
dmi_12         -8.0 (0.87)
aao_51         -3.1 (0.67)
olr_1           3.5 (0.79)
I(nino_4^2)     2.5 (0.54)
nino_4:olr_1    3.5 (0.76)
nino_4:dmi_12 -6.5 (0.77)
aao_51:olr_1   -2.3 (0.67)


Adjusted R-squared:  0.60
```

Smallest model highlights important climate-chemistry connections:

1. NINO has strong influence on CO at a four week lead time
2. Effect of DMI depends on length of lag

Interpretable models lead to scientific conclusions

$$\gamma = 1$$

	Est	(Std. Error)
(Intercept)	-1.6	(0.78)
nino_4	7.2	(0.78)
dmi_4	7.2	(0.93)
dmi_12	-8.0	(0.87)
aao_51	-3.1	(0.67)
olr_1	3.5	(0.79)
I(nino_4^2)	2.5	(0.54)
→ nino_4:olr_1	3.5	(0.76)
→ nino_4:dmi_12	-6.5	(0.77)
aao_51:olr_1	-2.3	(0.67)

Adjusted R-squared: 0.60

Smallest model highlights important climate-chemistry connections:

1. NINO has strong influence on CO at a four week lead time
2. Effect of DMI depends on length of lag
3. NINO interactions suggest that NINO amplifies effect of other indices

Model has good predictive skill

$$\gamma = 0$$

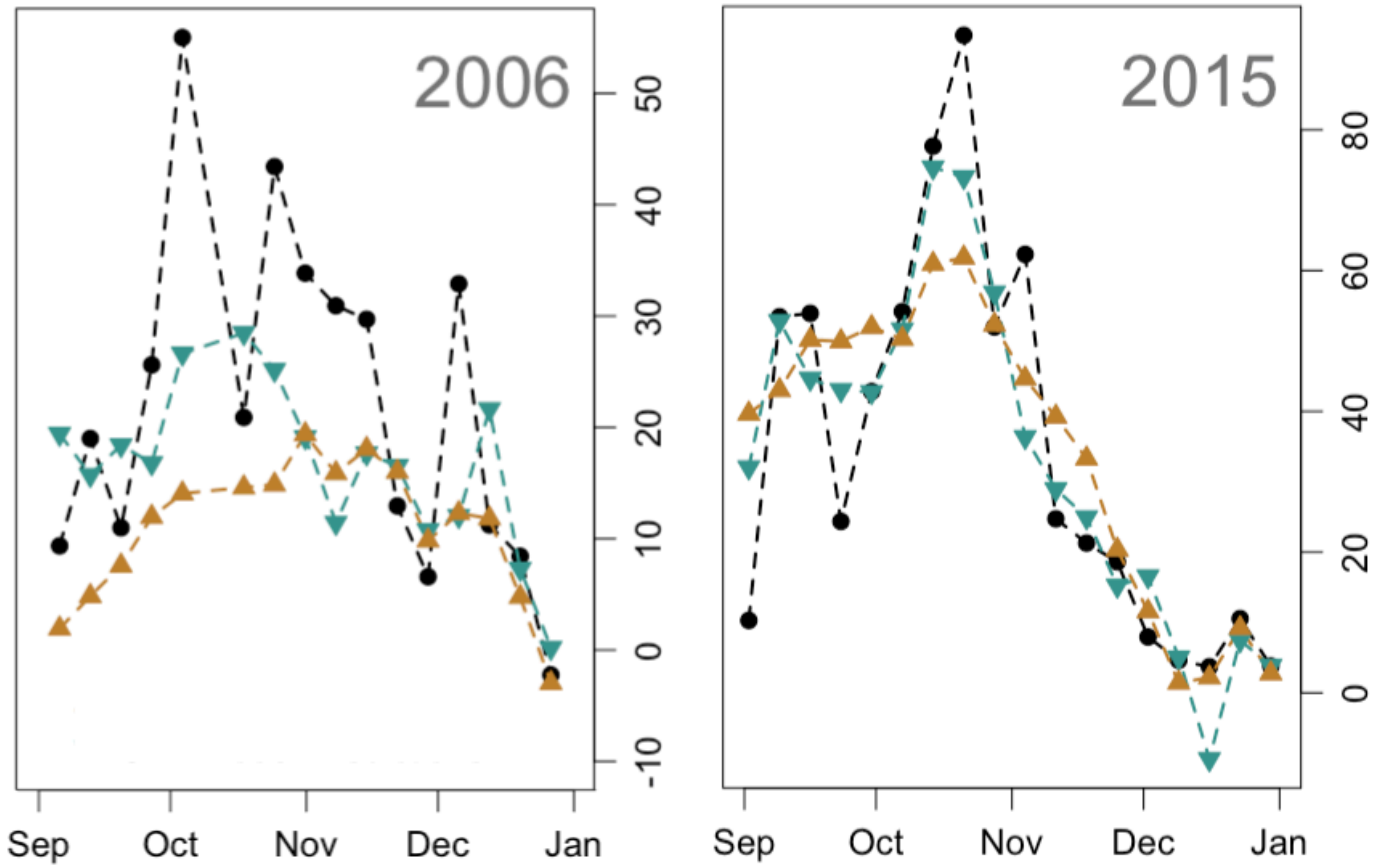
```
                Est (Std. Error)
(Intercept)          0.3 (0.70)
nino_4               7.6 (0.83)
dmi_1                5.7 (0.79)
dmi_12              -6.1 (0.75)
dmi_43               1.8 (0.65)
tsa_3               -2.2 (0.64)
aao_2               -3.6 (0.61)
aao_38              -2.2 (0.64)
aao_51              -1.6 (0.63)
olr_1                2.3 (0.74)
olr_13               3.4 (0.71)
nino_4:olr_1         3.2 (0.66)
nino_4:dmi_1         3.2 (0.81)
dmi_1:dmi_12        -4.5 (0.56)
nino_4:aao_51       -4.2 (0.77)
tsa_3:olr_1         -2.3 (0.63)
aao_2:olr_13        -2.1 (0.68)
nino_4:aao_2        -1.8 (0.70)

Adjusted R-squared:  0.68
```

OLR helps capture the most extreme CO anomalies

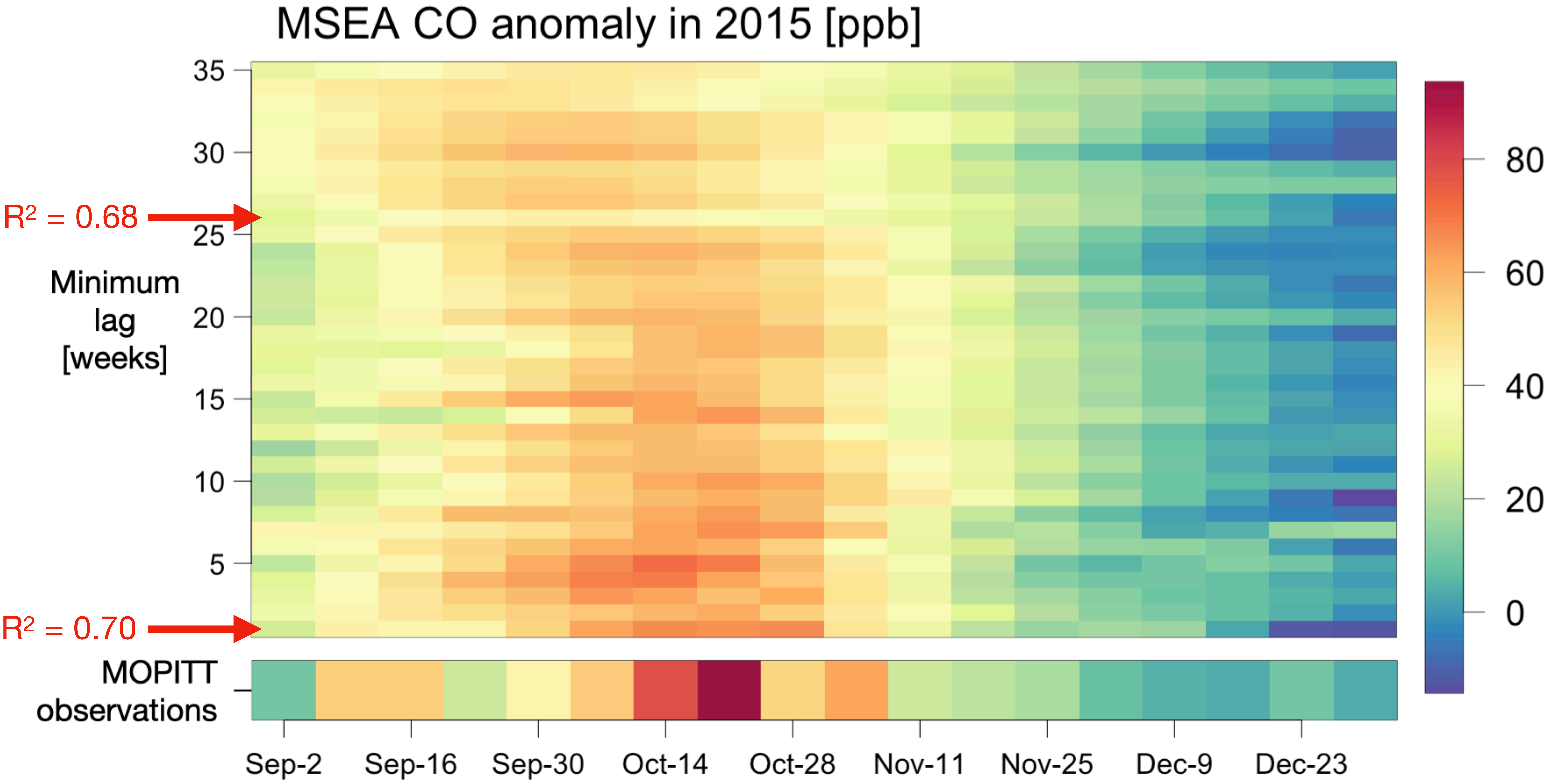
● Weekly Observations
▲ No OLR Model Predictions
▼ OLR Model Predictions

Adjusted R^2

No OLR Model	OLR Model
0.66	0.68



Model has good predictive skill at useful lead time

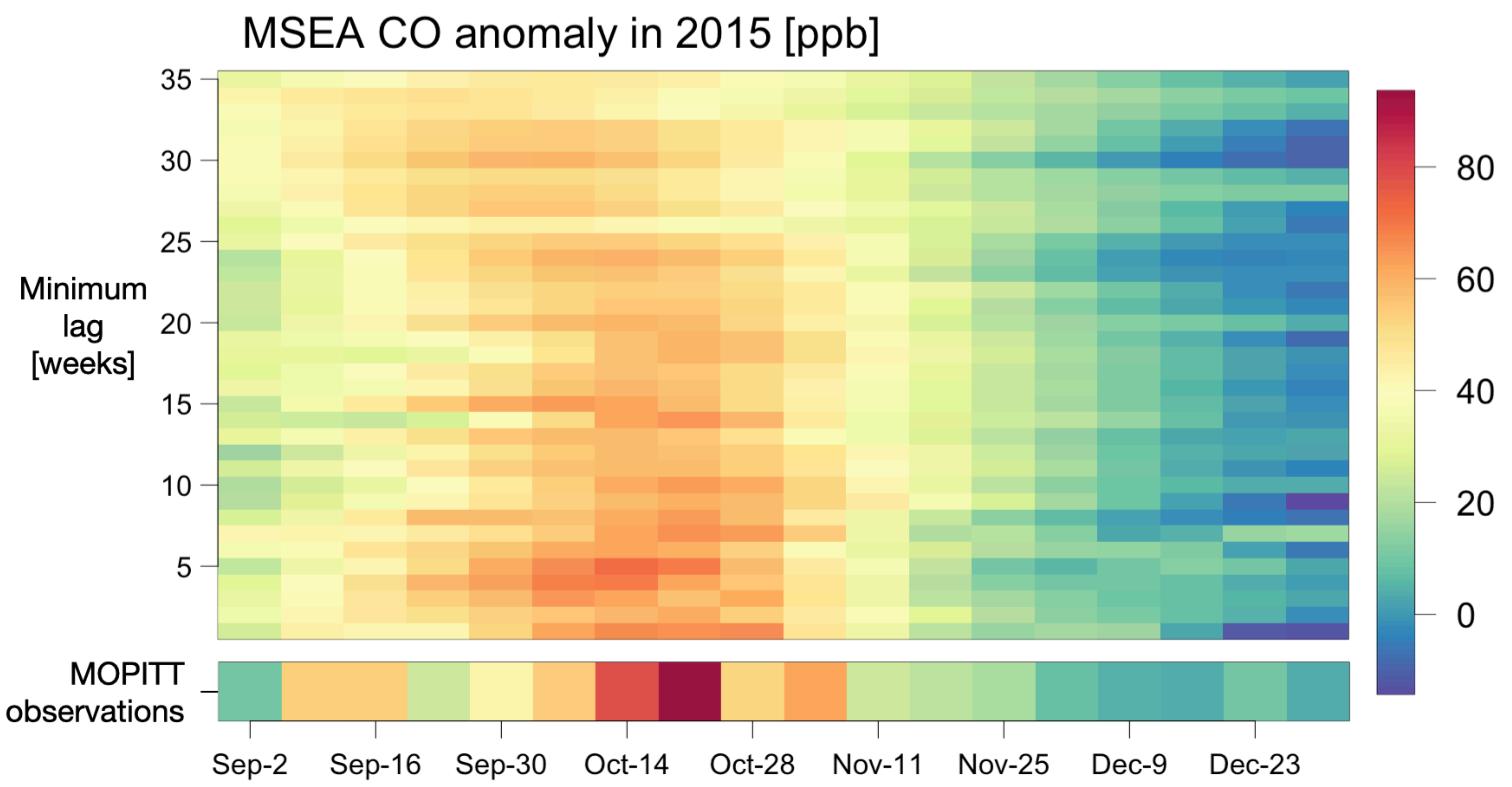


Conclusions

We are using natural variability in the climate to model atmospheric CO (a proxy for fire intensity)

- Interpretable models help explain natural drivers of fire season intensity
- Models have good predictive skill up to lead times of ~6 months in MSEA

```
             Est (Std. Error)
(Intercept)   -1.6 (0.78)
nino_4         7.2 (0.78)
dmi_4          7.2 (0.93)
dmi_12        -8.0 (0.87)
aao_51        -3.1 (0.67)
olr_1          3.5 (0.79)
I(nino_4^2)    2.5 (0.54)
nino_4:olr_1   3.5 (0.76)
nino_4:dmi_12 -6.5 (0.77)
aao_51:olr_1  -2.3 (0.67)


Adjusted R-squared:  0.60
```



Thank you! Questions?

Will Daniels

wsdaniels.github.io

wdaniels@mines.edu

See manuscript on EarthArXiv for details on research

See ISLR for details on stat learning methods