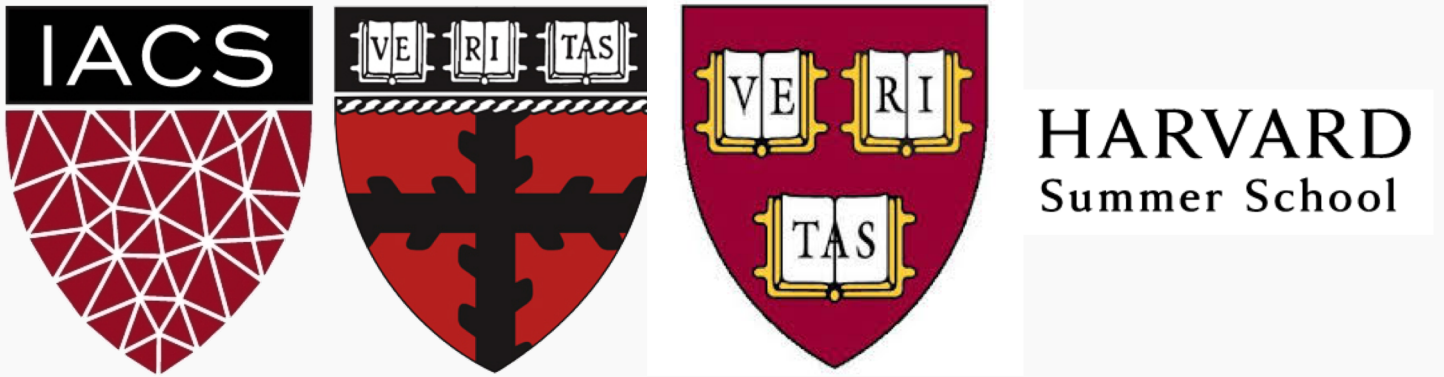
Lecture #4: Multiple Linear Regression

CS-S109A: Introduction to Data Science

Kevin Rader



ANNOUNCEMENTS

- **HW2** is due **Wed**, July 7 at 11:59pm
- **Lab2** is on Friday 12-2pm.
- **No Lecture** on Monday (happy 4th of July!).

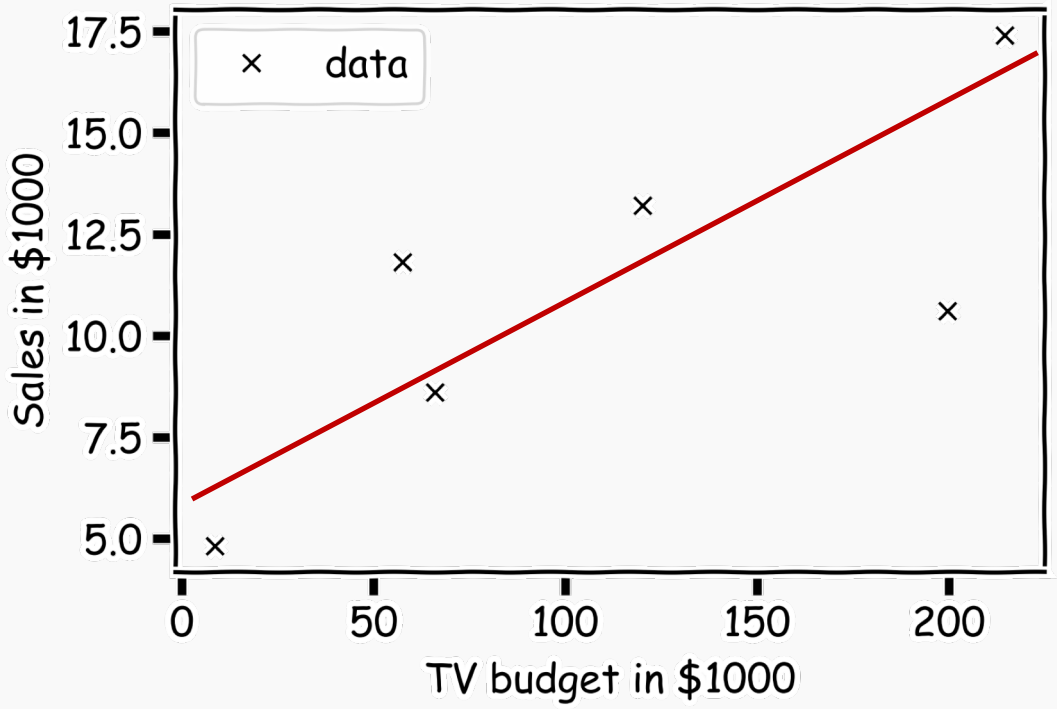
Lecture Outline

- Brief Recap
- How well do we know $\hat{f}$
  - Confidence and Prediction Intervals around our $\hat{f}$
- Multiple Regression
  - Formulation using Linear Algebra
- Categorical Predictors
- Interaction Terms
- Polynomial Regression
  - Linear Algebra Formulation
- Overfitting
- Variable Selection
- Bias vs. Variance

Summary from last lecture

We **assume** a simple form of the statistical model f:

$$Y = f(X) + \epsilon = \beta_0 + \beta_1 X + \epsilon$$

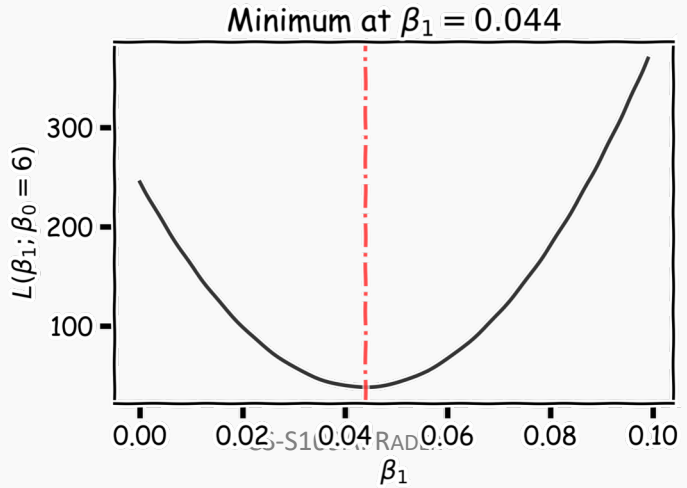


Summary from last lecture

We fit the model, i.e. estimate, $\hat{\beta}_0, \hat{\beta}_1$ that minimize the loss function, which we **assume** to be the MSE:

$$L_{MSE}(\beta_0, \beta_1) = \frac{1}{n}\sum_{n}[y_i - (\beta_0 + \beta_1 X)]^2$$

$$\widehat{\beta}_0, \widehat{\beta}_1 = \underset{\beta_0, \beta_1}{\operatorname{argmin}} L(\beta_0, \beta_1).$$



Summary from last lecture

We acknowledge that because there are errors in measurements and a limited sample, there is an inherent uncertainty in the estimation of $\hat{\beta}_0, \hat{\beta}_1$.

We used probability theory to determine the distributions of $\hat{\beta}_0, \hat{\beta}_1$

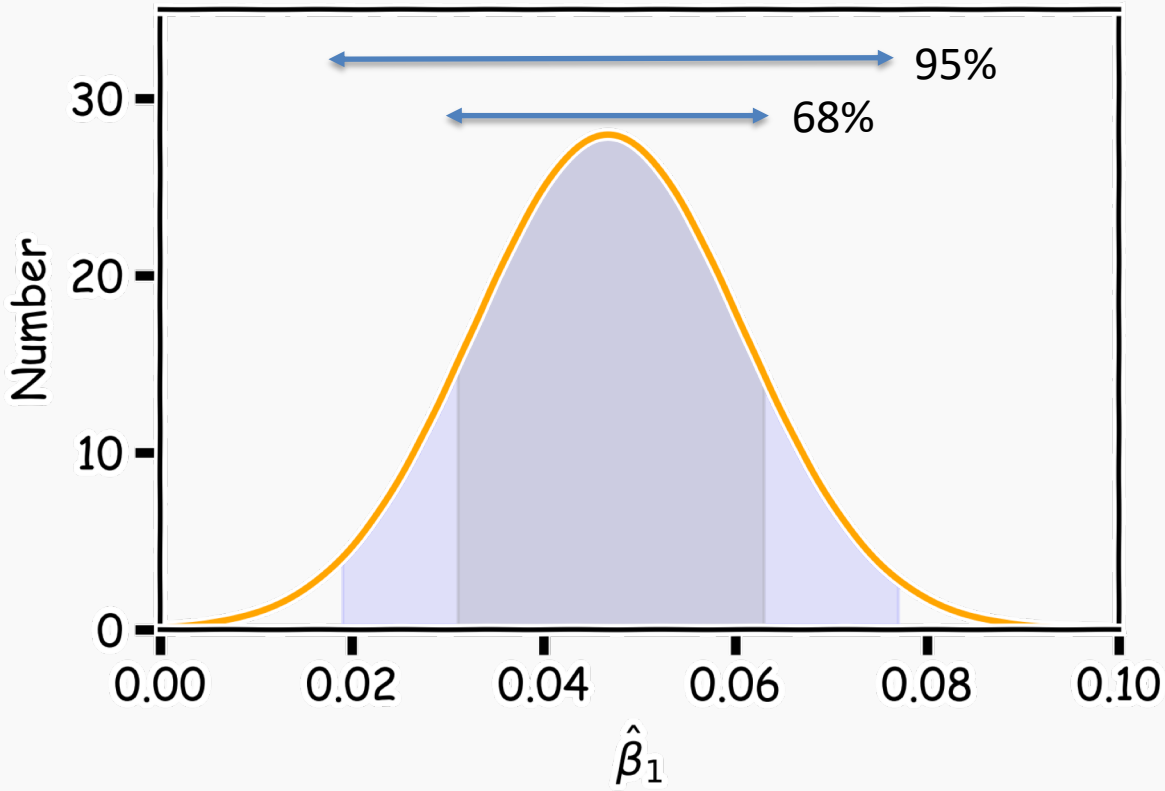
$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right)\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$$

*Note: σ^2 can be estimated (unbiased) with: $\hat{\sigma}^2 = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$.

Summary from last lecture

We calculate the confidence intervals, which are the ranges of values such that the **true** value of β_1 is contained in this interval with *n* percent of all such confidence interval that could be calculated. Practically speaking we build a Normal distribution around of $\hat{\beta}_1$ and pull off the relevant quantiles:



Summary from last lecture

We evaluate the importance of predictors using hypothesis testing, using *t*-statistics and *p*-values created assuming the null hypothesis is true:

To test the hypothesis $H_0: \beta_1 = 0$ vs. $H_0: \beta_1 \neq 0$ (in simple regression):

$$T = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2}}}$$

Use *p*-values or critical values from the *t*-distribution with $n - 2$ degrees of freedom.

How would this formula generalize if you would want to test a different value?

Alternative to Probability Theory: the Bootstrap

In the lack of active imagination, parallel universes and the likes, or trusting the probabilistic models, we need an alternative way of producing fake data set that resemble the parallel universes.

> Bootstrapping is the practice of sampling from the observed data (X, Y) in estimating statistical properties (sometimes called *resampling*).

For example, we can compute $\hat{\beta}_0$ and $\hat{\beta}_1$ multiple times by randomly resampling from our data set (the same sample size to maintain the same amount of uncertainty in each resample). We then use the variance of our multiple estimates to approximate the true variance of $\hat{\beta}_0$ and $\hat{\beta}_1$.

Lecture Outline

Brief Recap

How well do we know $\hat{f}$

Confidence and Prediction Intervals around our $\hat{f}$

Multiple Regression

Formulation using Linear Algebra

Categorical Predictors

Interaction Terms

Polynomial Regression

Linear Algebra Formulation

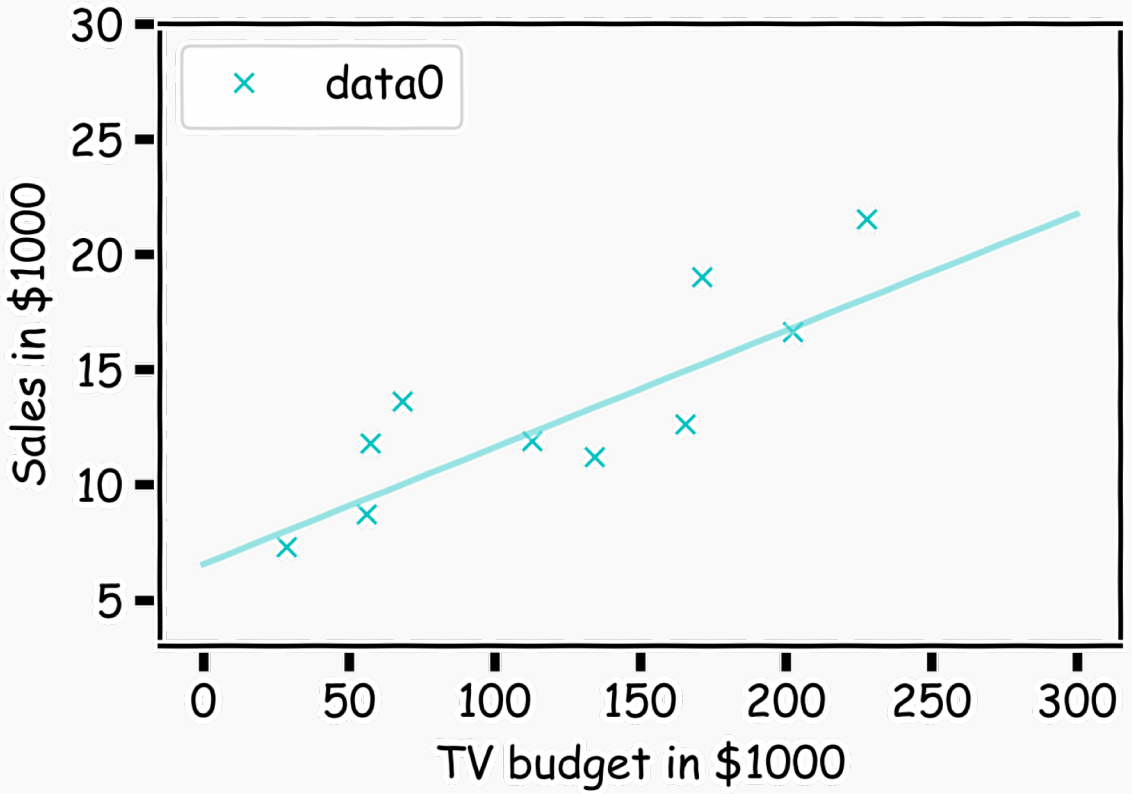
Overfitting

Variable Selection

Bias vs. Variance

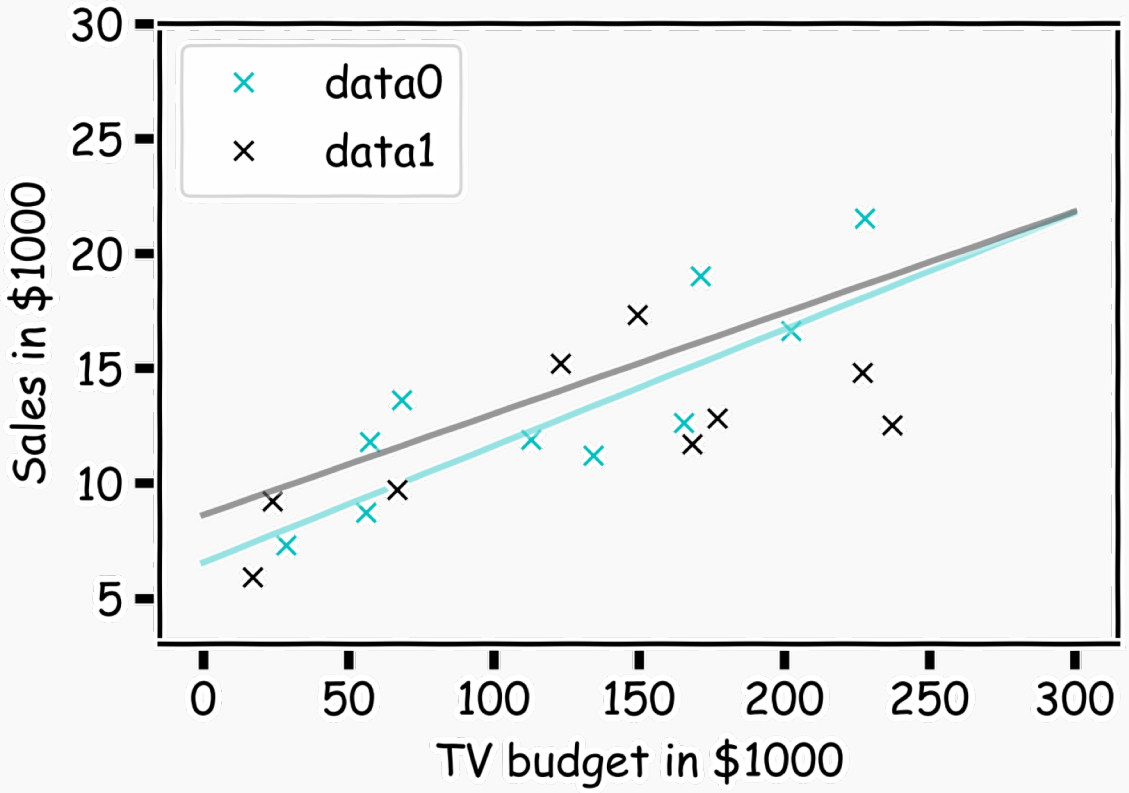
How well do we know $\hat{f}$?

Our confidence in f is directly connected with the confidence in βs. So for each bootstrap sample, we have one β_0, β_1 which we can use to predict y for all x's.



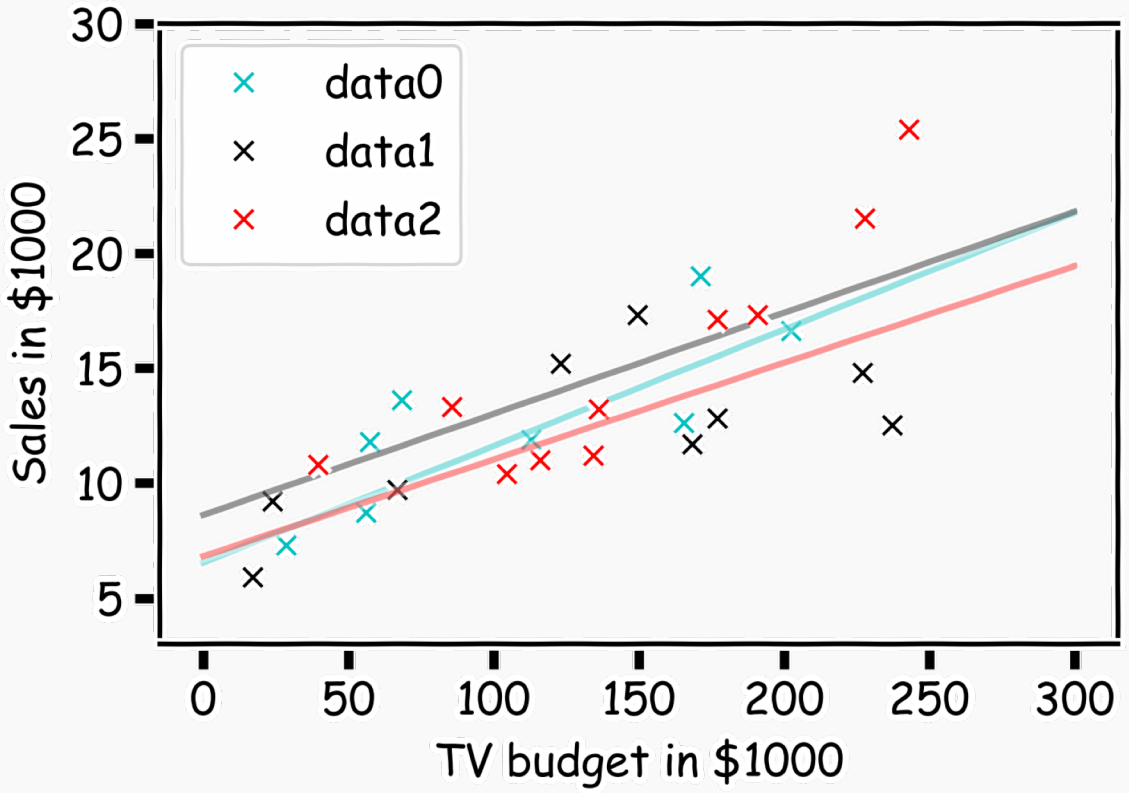
How well do we know $\hat{f}$?

Here we show two difference set of models given the fitted coefficients.



How well do we know $\hat{f}$?

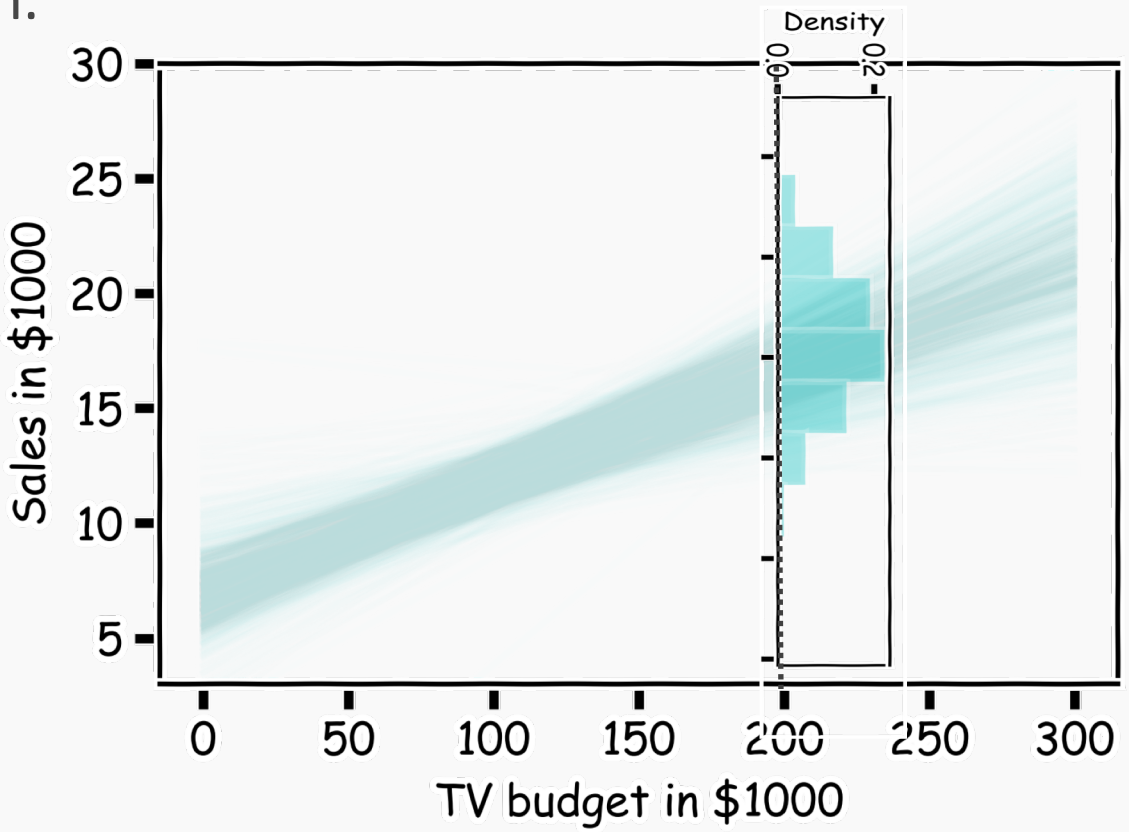
There is one such regression line for every bootstrapped sample.



How well do we know $\hat{f}$?

Below we show all regression lines for a thousand of such bootstrapped samples.

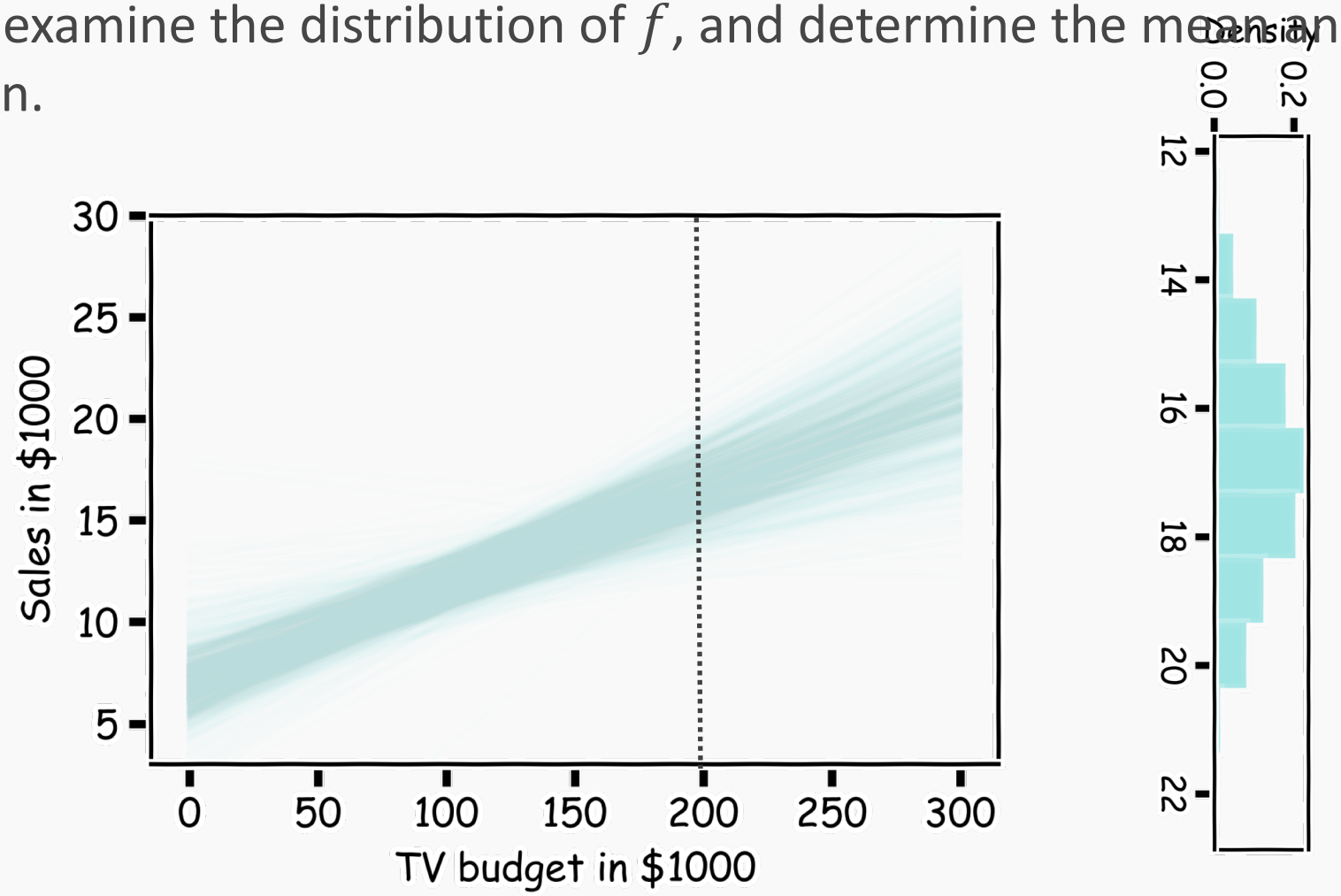
For a given x, we examine the distribution of $\hat{f}$, and determine the mean and standard deviation.



How well do we know $\hat{f}$?

Below we show all regression lines for a thousand of such sub-samples.

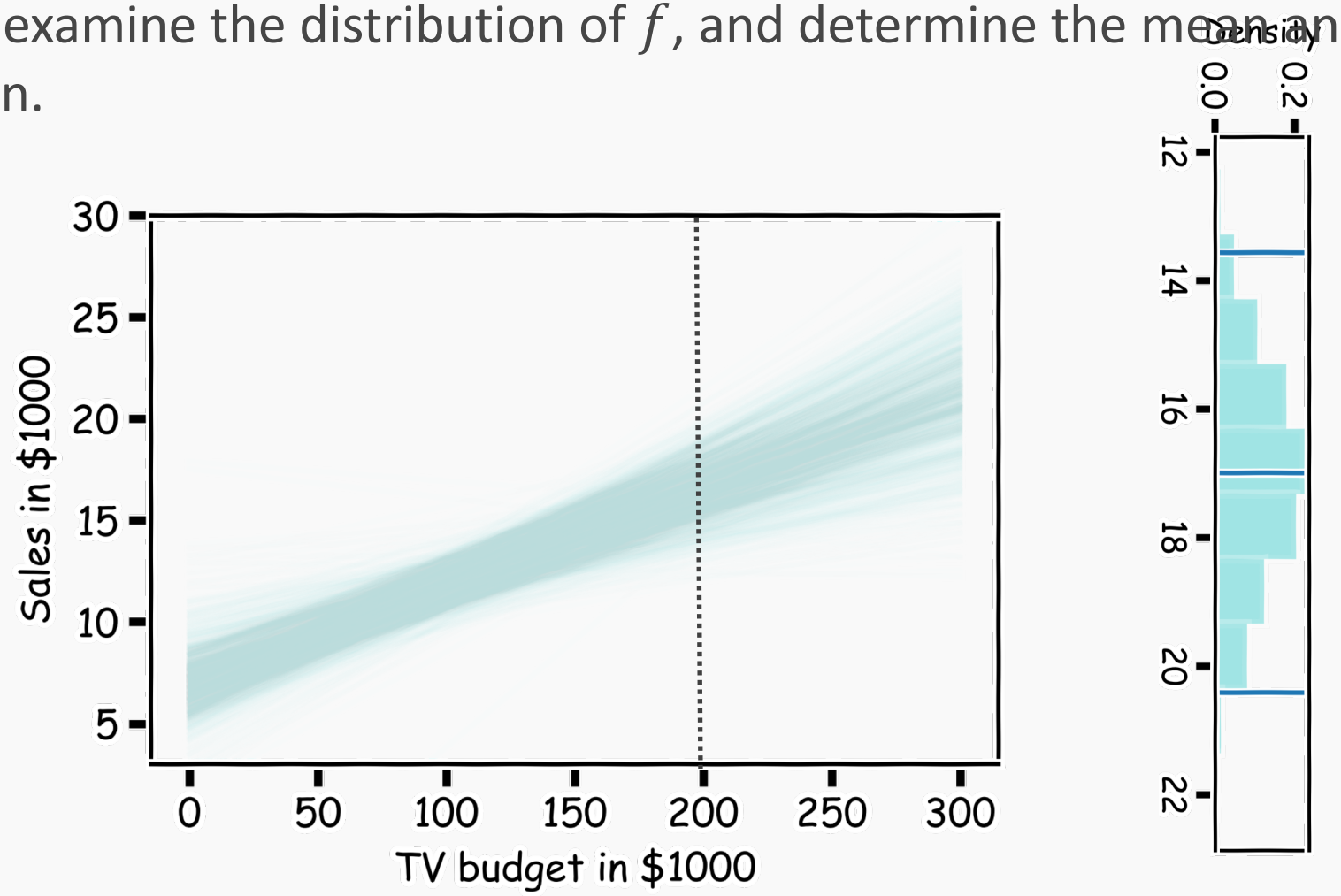
For a given x, we examine the distribution of $\hat{f}$, and determine the mean and standard deviation.



How well do we know $\hat{f}$?

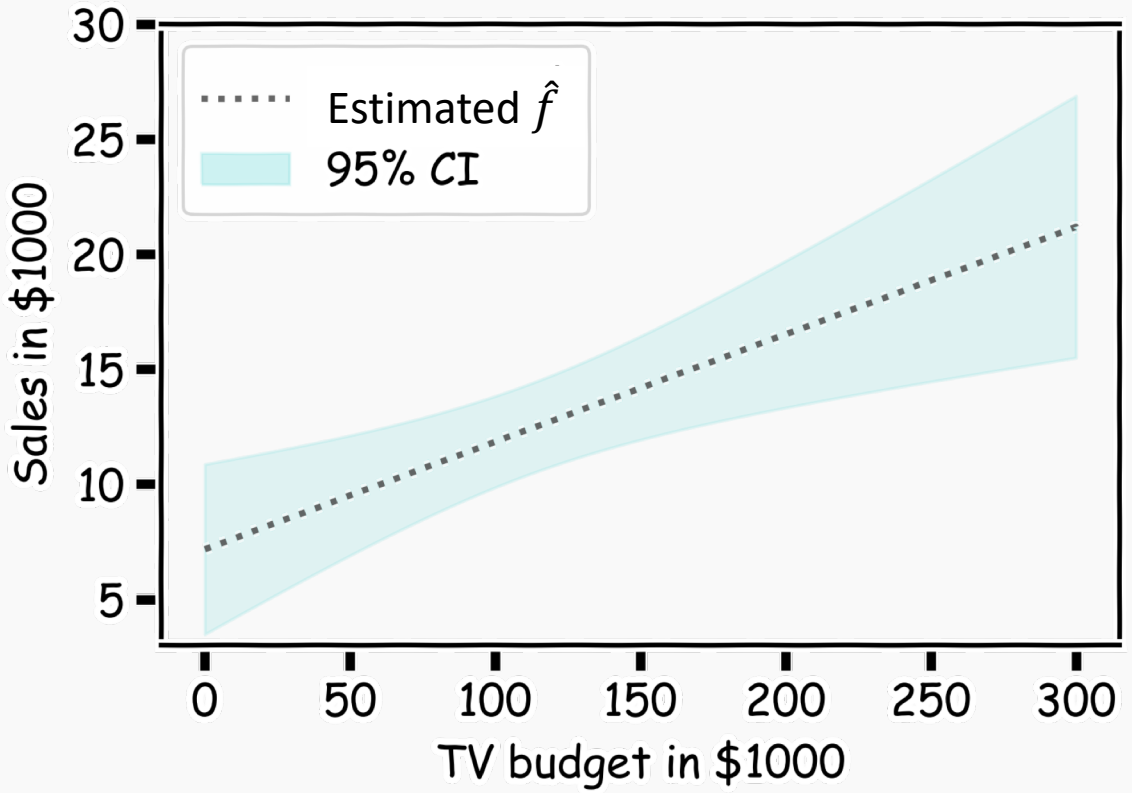
Below we show all regression lines for a thousand of such sub-samples.

For a given x, we examine the distribution of $\hat{f}$, and determine the mean and standard deviation.

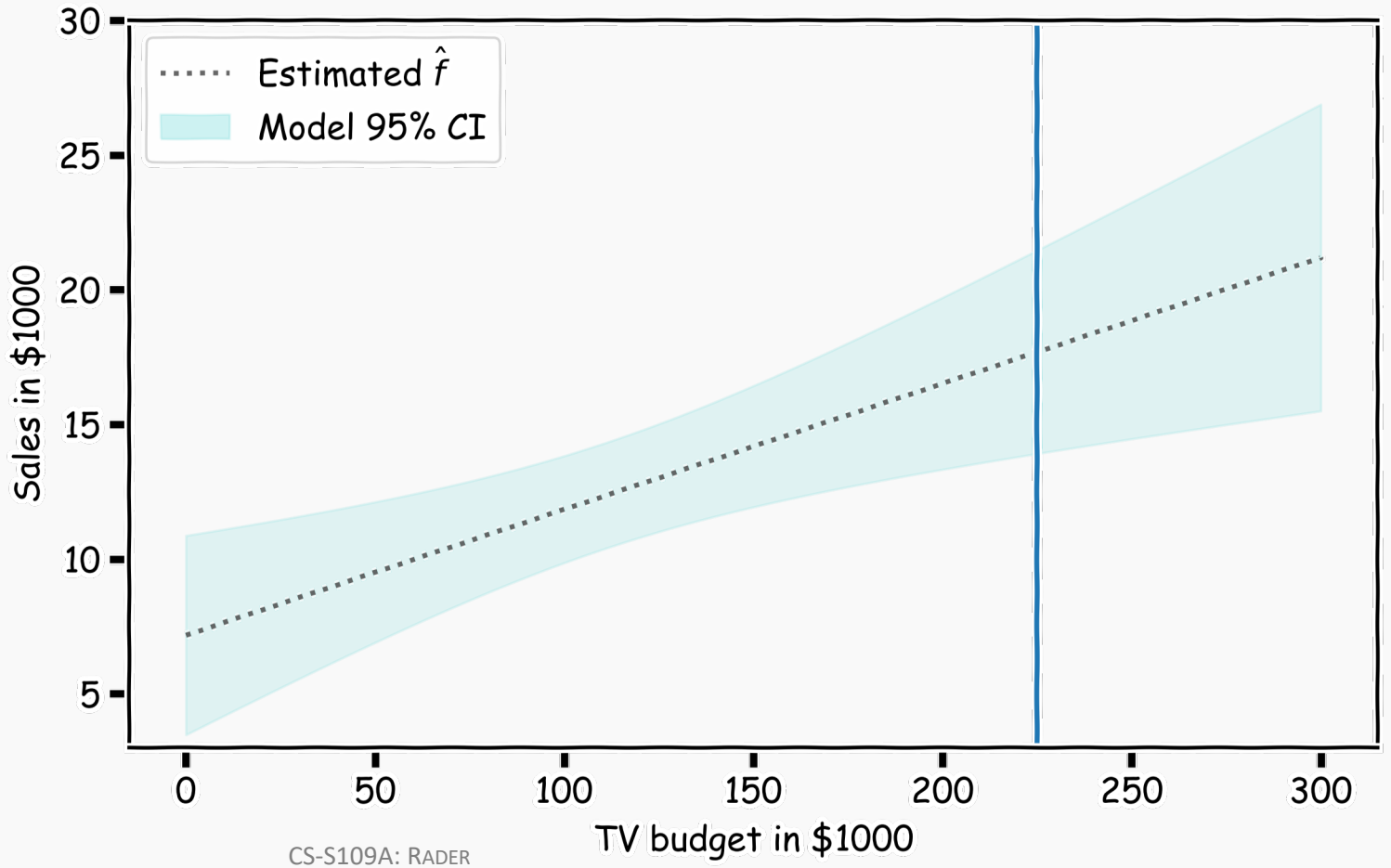


How well do we know $\hat{f}$?

For every x, we calculate the mean of the models, $\hat{f}$ (shown with dotted line) and the 95% CI of those models (shaded area).

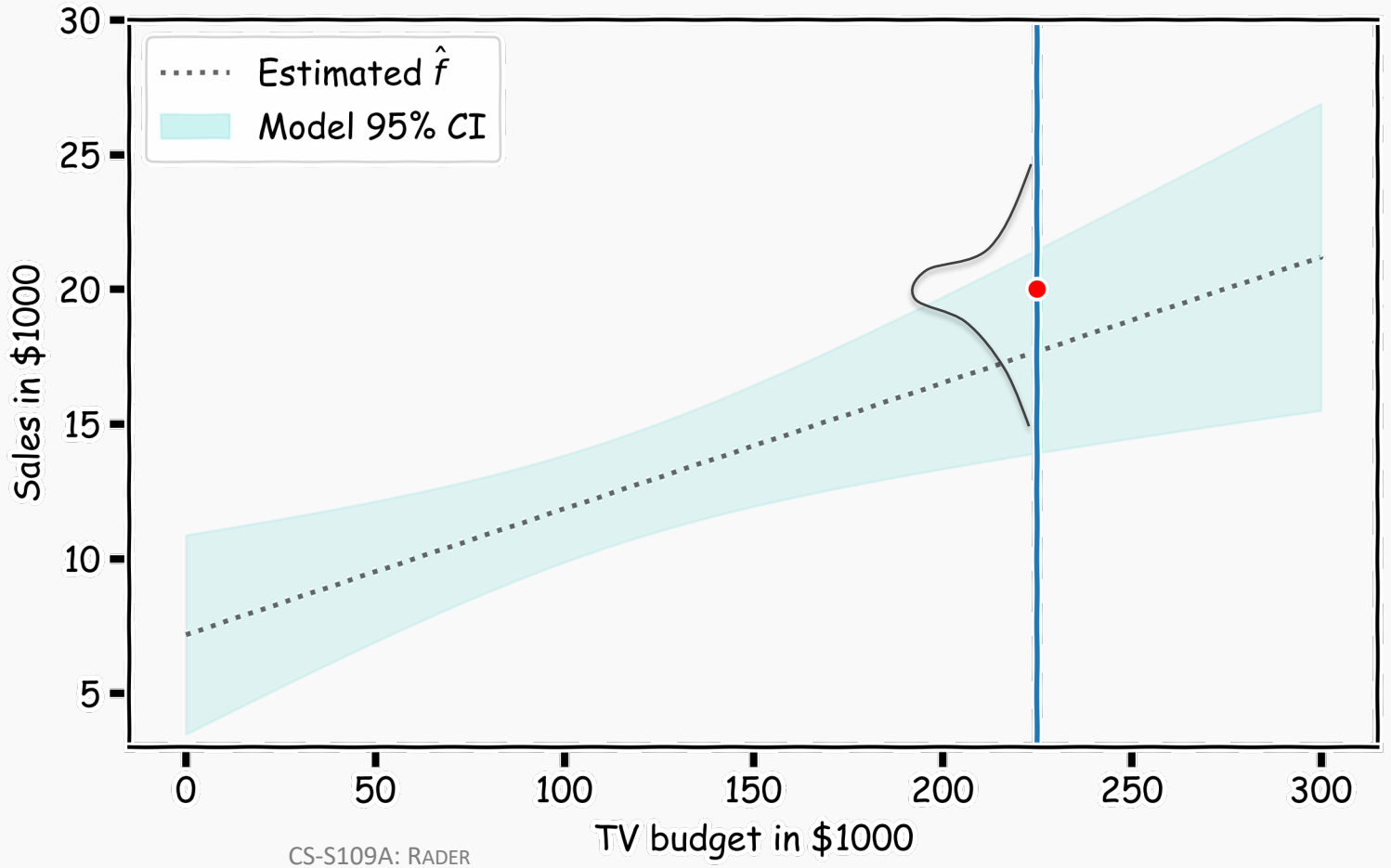


Confidence in predicting $\hat{y}$



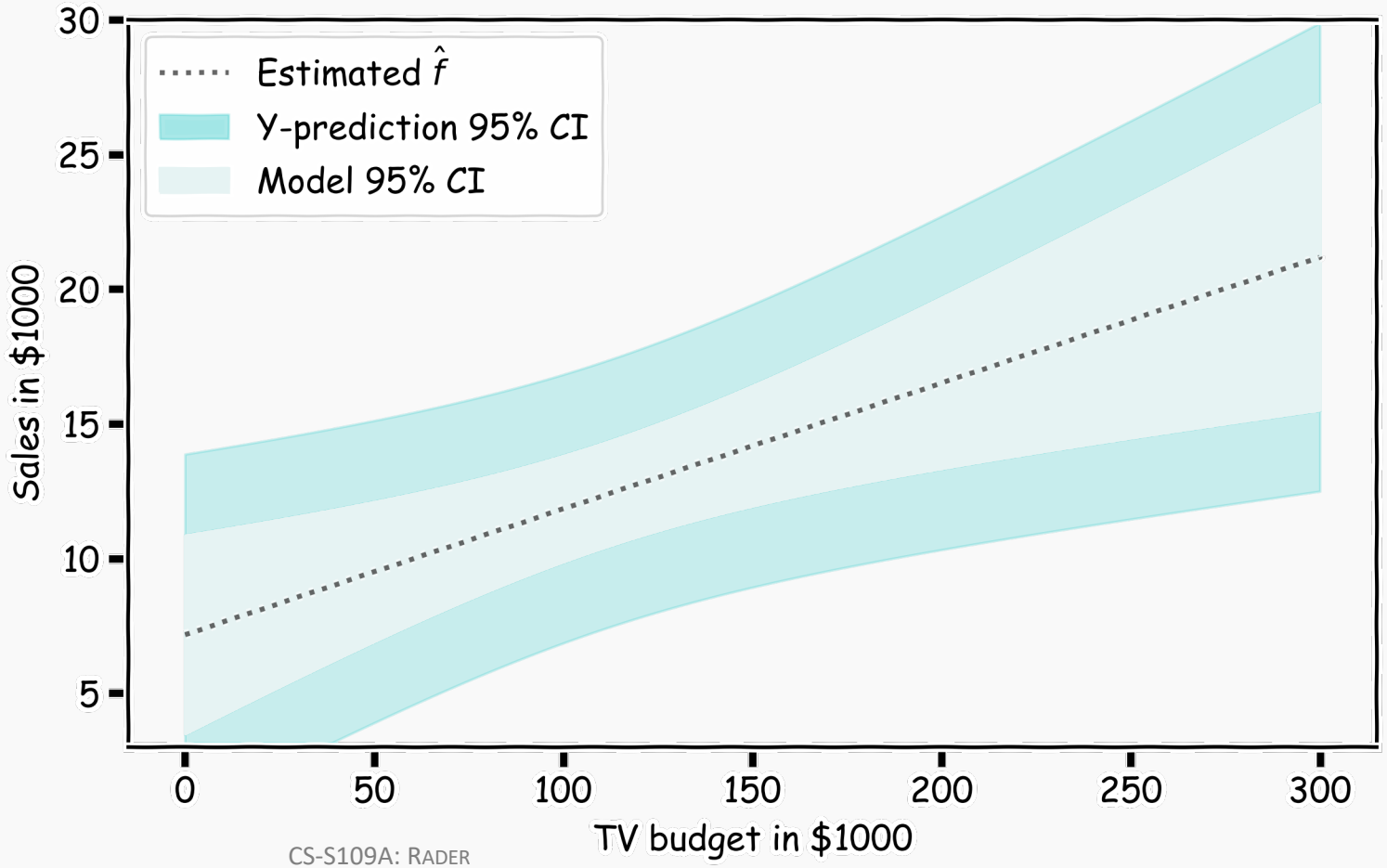
Confidence in predicting $\hat{y}$

- for a given x, we have a distribution of models $f(x)$
- for each of these $f(x)$, the prediction for $y \sim N(f, \sigma_\epsilon)$



Confidence in predicting $\hat{y}$

- for a given x, we have a distribution of models $f(x)$
- for each of these $f(x)$, the prediction for $y \sim N(f, \sigma_\epsilon)$
- The prediction confidence intervals are then



Lecture Outline

- Brief Recap
- How well do we know $\hat{f}$
  - Confidence and Prediction Intervals around our $\hat{f}$
- **Multiple Regression**
  - **Formulation using Linear Algebra**
- Categorical Predictors
- Interaction Terms
- Polynomial Regression
  - Linear Algebra Formulation
- Overfitting
- Variable Selection
- Bias vs. Variance

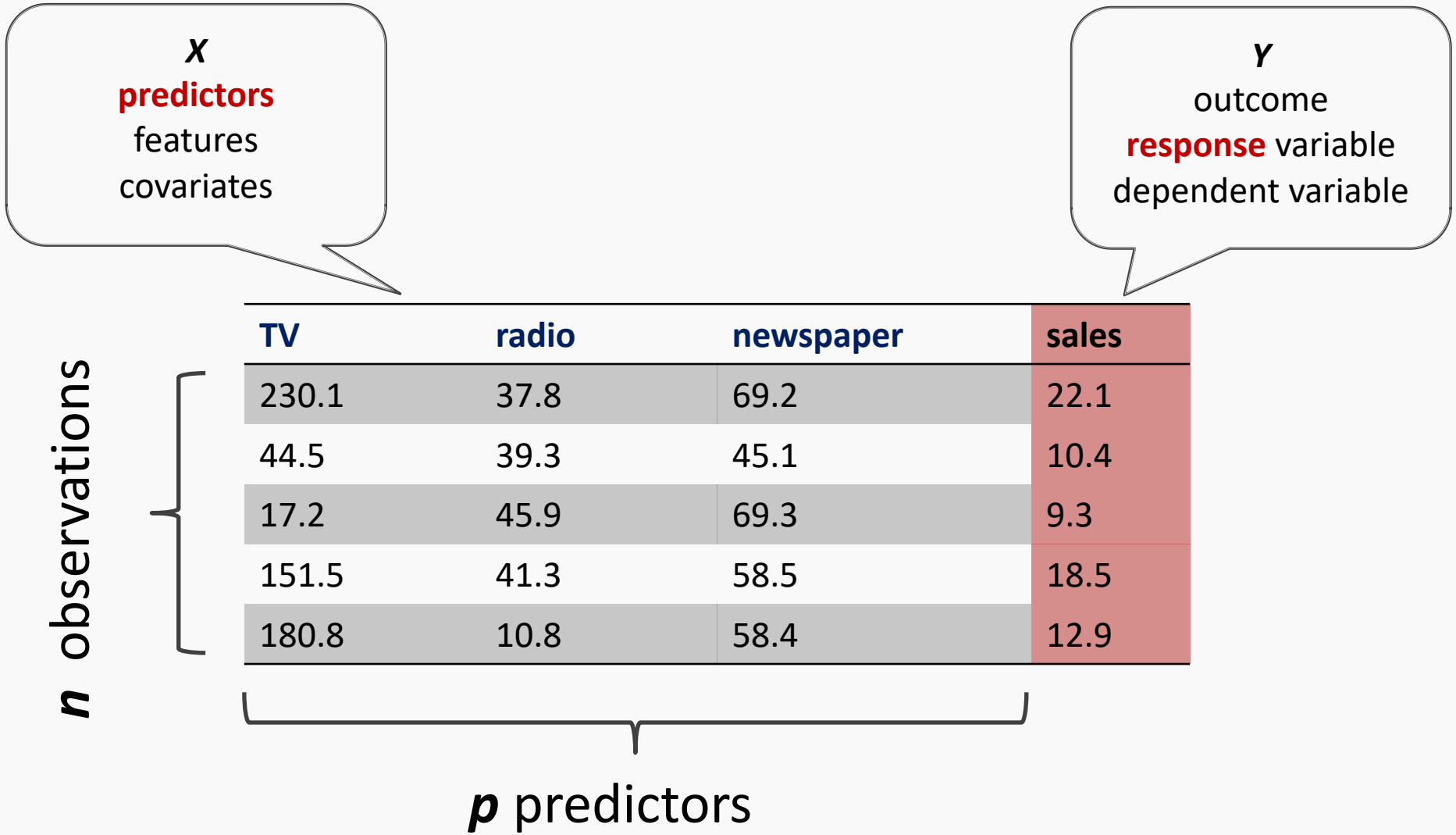
Multiple Linear Regression

If you have to guess someone's height, would you rather be told

- Their age only
- Their age and gender
- Their age, gender, and income
- Their age, gender, income, and favorite number

Of course, you'd always want as much data about a person as possible. Even though height and favorite number may not be strongly related, at worst you could just ignore the information on favorite number. We want our models to be able to take in lots of data as they make their predictions.

Response vs. Predictor Variables



TV	radio	newspaper	sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

Linear Models with Multiple Predictors

In practice, it is unlikely that any response variable Y depends solely on one predictor *x*. Rather, we expect that is a function of multiple predictors $f(X_1, \ldots, X_J)$. Using the notation we introduced last lecture,

$$Y = y_1, \ldots, y_n, \quad X = X_1, \ldots, X_J \text{ and } X_j = x_{1j}, \ldots, x_{ij}, \ldots, x_{nj}$$

In this case, we can still assume a simple form for f -a multilinear form:

$$Y = f(X_1, \ldots, X_J) + \epsilon = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_J X_J + \epsilon$$

Hence, $\hat{f}$, has the form

$$\hat{Y} = \hat{f}(X_1, \ldots, X_J) + \epsilon = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_J X_J + \epsilon$$

Multiple Linear Regression

Again, to fit this model means to compute $\hat{\beta}_0, \ldots, \hat{\beta}_J$ or to minimize a loss function; we will again choose the **MSE** as our loss function.

Given a set of observations,

$$\{(x_{1,1}, \ldots, x_{1,J}, y_1), \ldots (x_{n,1}, \ldots, x_{n,J}, y_n)\},$$

the data and the model can be expressed in vector notation,

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_y \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,J} \\ 1 & x_{2,1} & \ldots & x_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \ldots & x_{n,J} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_J \end{pmatrix},$$

Multilinear Model, example

For our data

$$\text{Sales} = \beta_0 + \beta_1 \times TV + \beta_2 \times Radio + \beta_3 \times Newspaper + \epsilon$$

In linear algebra notation

$$\boldsymbol{Y} = \begin{pmatrix} Sales_1 \\ \vdots \\ Sales_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & TV_1 & Radio_1 & News_1 \\ \vdots & \vdots & \vdots \\ 1 & TV_n. & Radio_n & News_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_3 \end{pmatrix}$$

$$\begin{pmatrix} Sales_1 \end{pmatrix} = \begin{pmatrix} 1 & TV_1 & Radio_1 & News_1 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_3 \end{pmatrix}$$

Multiple Linear Regression

The model takes a simple algebraic form:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

Thus, the MSE can be expressed in vector notation as

$$\text{MSE}(\beta) = \frac{1}{\text{n}}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$$

Minimizing the MSE using vector calculus yields,

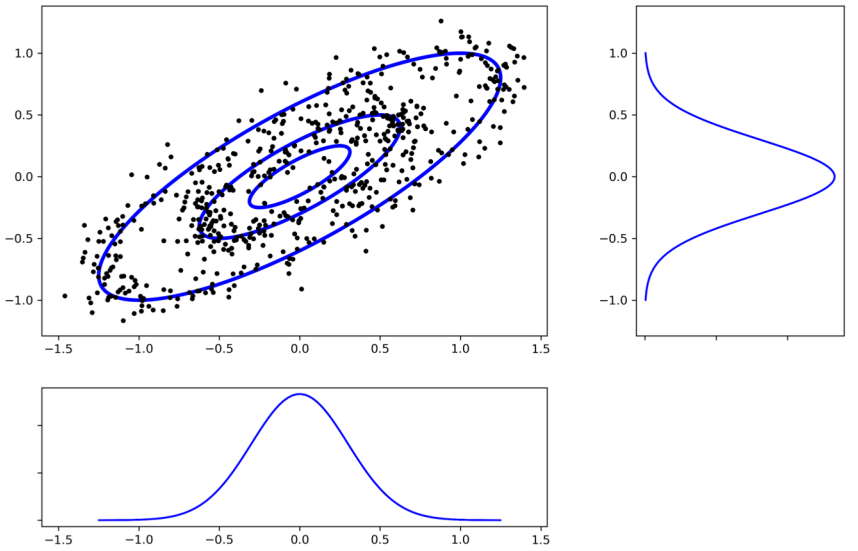
$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y} = \underset{\boldsymbol{\beta}}{\text{argmin}}\ \text{MSE}(\boldsymbol{\beta}).$$

Standard Errors for Multiple Linear Regression

As with the simple linear regression, the standard errors can be calculated either using statistical modeling

$$\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

Alternatively, use the **bootstrap.**



Collinearity

Collinearity (sometimes called multicollinearity) refers to the case in which two or more predictors are correlated (related).

We will re-visit collinearity in the next lecture when we address **overfitting**, but for now we want to examine how does collinearity affects our confidence on the coefficients and consequently on the importance of those coefficients.

Collinearity

Three individual models

TV

Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
6.679	0.478	13.957	2.804e-31	5.735	7.622
0.048	0.0027	17.303	1.802e-41	0.042	0.053

RADIO

Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
9.567	0.553	17.279	2.133e-41	8.475	10.659
0.195	0.020	9.429	1.134e-17	0.154	0.236

NEWS

Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
11.55	0.576	20.036	1.628e-49	10.414	12.688
0.074	0.014	5.134	6.734e-07	0.0456	0.102

One model

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
β_0	2.602	0.332	7.820	3.176e-13	1.945	3.258
β_{TV}	0.046	0.0015	29.887	6.314e-75	0.043	0.049
β_{RADIO}	0.175	0.0094	18.576	4.297e-45	0.156	0.194
β_{NEWS}	0.013	0.028	2.338	0.0203	0.008	0.035

Finding Significant Predictors: Hypothesis Testing

For checking the significance of linear regression coefficients:

1.we set up our hypotheses H_0:

$$H_0 : \beta_0 = \beta_1 = \ldots = \beta_J = 0 \qquad \textbf{(Null)}$$

$$H_1 : \beta_j \neq 0, \text{ for at least one } j \qquad \textbf{(Alternative)}$$

2. we choose the *F*-stat to evaluate the null hypothesis,

$$F = \frac{\text{explained variance}}{\text{unexplained variance}}$$

Finding Significant Predictors: Hypothesis Testing

3. we can compute the *F*-stat for linear regression models by

$$F = \frac{(\text{TSS} - \text{RSS})/J}{\text{RSS}/(n - J - 1)}, \quad \text{TSS} = \sum_i (y_i - \overline{y})^2, \text{RSS} = \sum_i (y_i - \widehat{y}_i)^2$$

4. If $F \approx 1$ we consider this evidence for H_0; if $F >> 1$, we consider this evidence against H_0.

Lecture Outline

- Brief Recap
- How well do we know $\hat{f}$
  - Confidence and Prediction Intervals around our $\hat{f}$
- Multiple Regression
  - Formulation using Linear Algebra
- **Categorical Predictors**
- Interaction Terms
- Polynomial Regression
  - Linear Algebra Formulation
- Overfitting
- Variable Selection
- Bias vs. Variance

Qualitative Predictors

So far, we have assumed that all variables are quantitative. But in practice, often some predictors are **qualitative**.

Example: The Credit data set contains information about balance, age, cards, education, income, limit , and rating for a number of potential customers.

Income	Limit	Rating	Cards	Age	Education	Gender	Student	Married	Ethnicity	Balance
14.890	3606	283	2	34	11	Male	No	Yes	Caucasian	333
106.02	6645	483	3	82	15	Female	Yes	Yes	Asian	903
104.59	7075	514	4	71	11	Male	No	No	Asian	580
148.92	9504	681	3	36	11	Female	No	No	Asian	964
55.882	4897	357	2	68	16	Male	No	Yes	Caucasian	331

Qualitative Predictors

If the predictor takes only two values, then we create an **indicator** or **dummy variable** that takes on two possible numerical values.

For example for the gender, we create a new variable:

$$x_i = \begin{cases} 1 & \text{if } i\,\text{th person is female} \\ 0 & \text{if } i\,\text{th person is male} \end{cases}$$

We then use this variable as a predictor in the regression equation.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\,\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\,\text{th person is male} \end{cases}$$

Qualitative Predictors

Question: What is interpretation of β_0 and β_1?

Qualitative Predictors

Question: What is interpretation of β_0 and β_1?

- β_0 is the average credit card balance among males,
- $\beta_0 + \beta_1$ is the average credit card balance among females,
- and β_1 the average difference in credit card balance between females and males.

Example: Calculate β_0 and β_1 for the Credit data.

You should find $\beta_0 \sim \$509, \beta_1 \sim \$19$

More than two levels: One hot encoding

Often, the qualitative predictor takes more than two values (e.g. ethnicity in the credit data).

In this situation, a single dummy variable cannot represent all possible values.

We create additional dummy variable as:

$$x_{i,1} = \begin{cases} 1 & \text{if } i\,\text{th person is Asian} \\ 0 & \text{if } i\,\text{th person is not Asian} \end{cases}$$

$$x_{i,2} = \begin{cases} 1 & \text{if } i\,\text{th person is Caucasian} \\ 0 & \text{if } i\,\text{th person is not Caucasian} \end{cases}$$

More than two levels: One hot encoding

We then use these variables as predictors, the regression equation becomes:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\,\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\,\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\,\text{th person is AfricanAmerican} \end{cases}$$

Question: What is the interpretation of β_0, β_1, β_2?

Lecture Outline

- Brief Recap
- How well do we know $\hat{f}$
  - Confidence and Prediction Intervals around our $\hat{f}$
- Multiple Regression
  - Formulation using Linear Algebra
- Categorical Predictors
- **Interaction Terms**
- Polynomial Regression
  - Linear Algebra Formulation
- Overfitting
- Variable Selection
- Bias vs. Variance

Beyond linearity

In the Advertising data, we assumed that the effect on sales of increasing one advertising medium is independent of the amount spent on the other media.

If we assume linear model then the average effect on sales of a one-unit increase in TV is always β_1, regardless of the amount spent on radio.

Synergy effects or **interaction effects** apply when an increase on the radio budget affects the association of TV spending and sales.

Beyond linearity

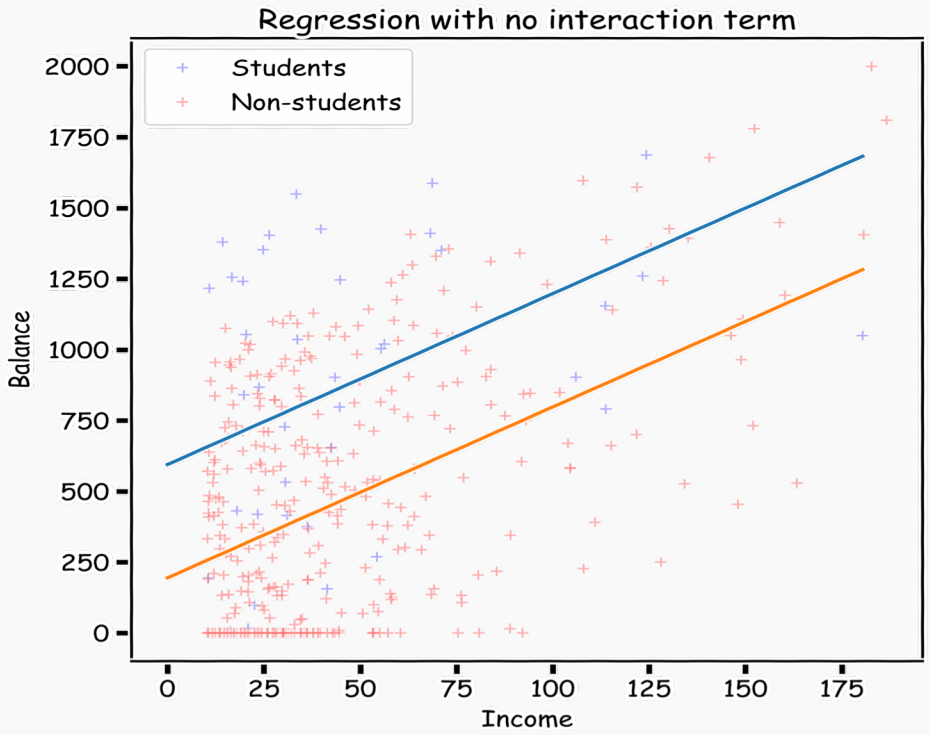
We change

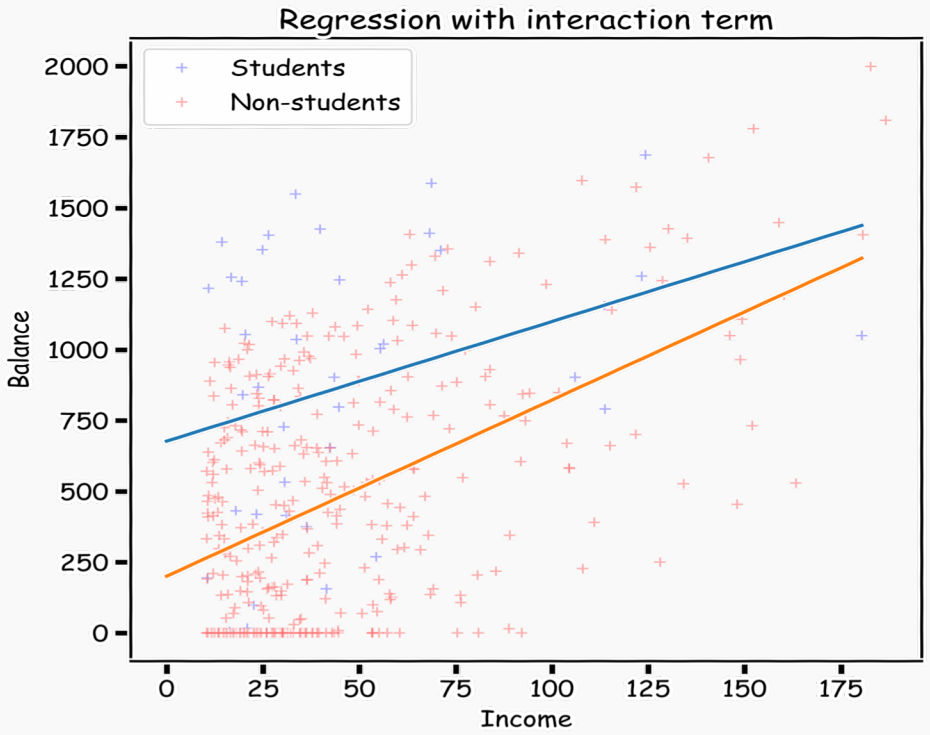
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

To

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

What does it mean?





$$x_{Student} = \begin{cases} 0 & Balance = \beta_0 + \beta_1 \times Income. \\ 1 & Balance = (\beta_0 + \beta_2) + (\beta_1) \times Income. \end{cases}$$

$$x_{Student} = \begin{cases} 0 & Balance = \beta_0 + \beta_1 \times Income. \\ 1 & Balance = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times Income \end{cases}$$

Predictors predictors predictors

We have a lot predictors!

Is it a problem?

Yes: Computational Cost

Yes: Overfitting

Wait there is more …

Residuals

We started with

$$y = f(x) + \epsilon$$

We **assumed** the exact form of $f(x)$,to be,

$$f(x) = \beta_0 + \beta_1 x,$$

then estimated the $\hat{\beta}'s$.

What if that is not correct? Instead:

$$f(x) = \beta_0 + \beta_1 x + \phi(x),$$

But we model it as

$$\hat{y} = \hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

Then the residual

$$r = (y - \widehat{y}) = \hat{f}(x) = \epsilon + \phi(x)$$

Residuals

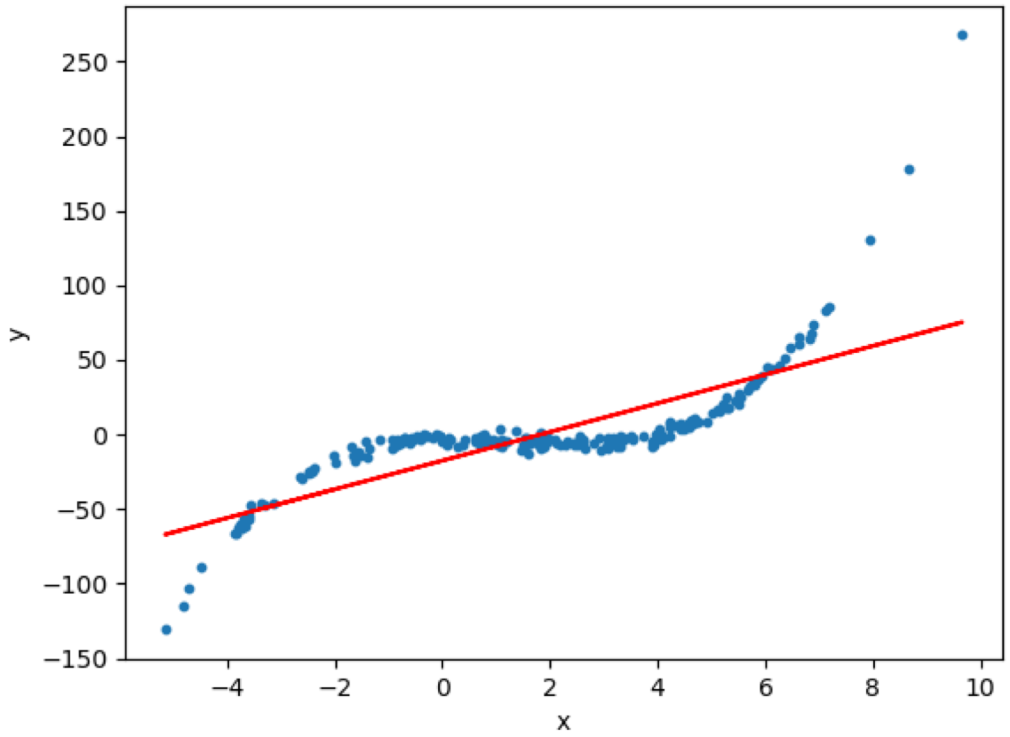
Residual Analysis

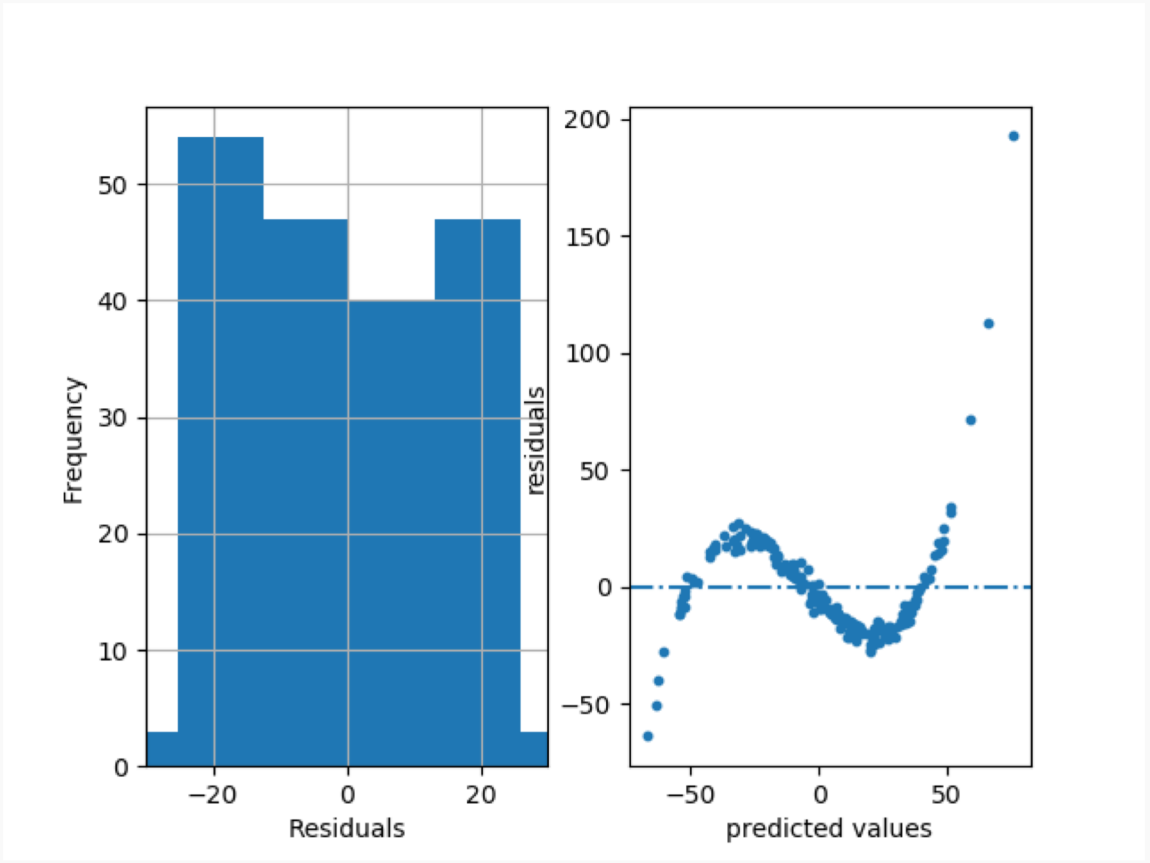
When we estimated the variance of ϵ, we assumed that the residuals $r_i = y_i - \hat{y}_i$ were uncorrelated and normally distributed with mean 0 and fixed variance.

These assumptions need to be verified using the data. In residual analysis, we typically create two types of plots:

1. a plot of r_i with respect to x_i or $\hat{y}_i$. This allows us to compare the distribution of the noise at different values of x_i.
2. a histogram of r_i. This allows us to explore the distribution of the noise independent of x_i or $\hat{y}_i$.

Residual Analysis





Lecture Outline

- Brief Recap
- How well do we know $\hat{f}$
  - Confidence and Prediction Intervals around our $\hat{f}$
- Multiple Regression
  - Formulation using Linear Algebra
- Categorical Predictors
- Interaction Terms
- **Polynomial Regression**
  - **Linear Algebra Formulation**
- Overfitting
- Variable Selection
- Bias vs. Variance

Polynomial Regression

The simplest non-linear model we can consider, for a response *Y* and a predictor *X*, is a polynomial model of degree *M*,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_M x^M + \epsilon.$$

Just as in the case of linear regression with cross terms, polynomial regression is a special case of linear regression - we treat each x^m as a separate predictor. Thus, we can write

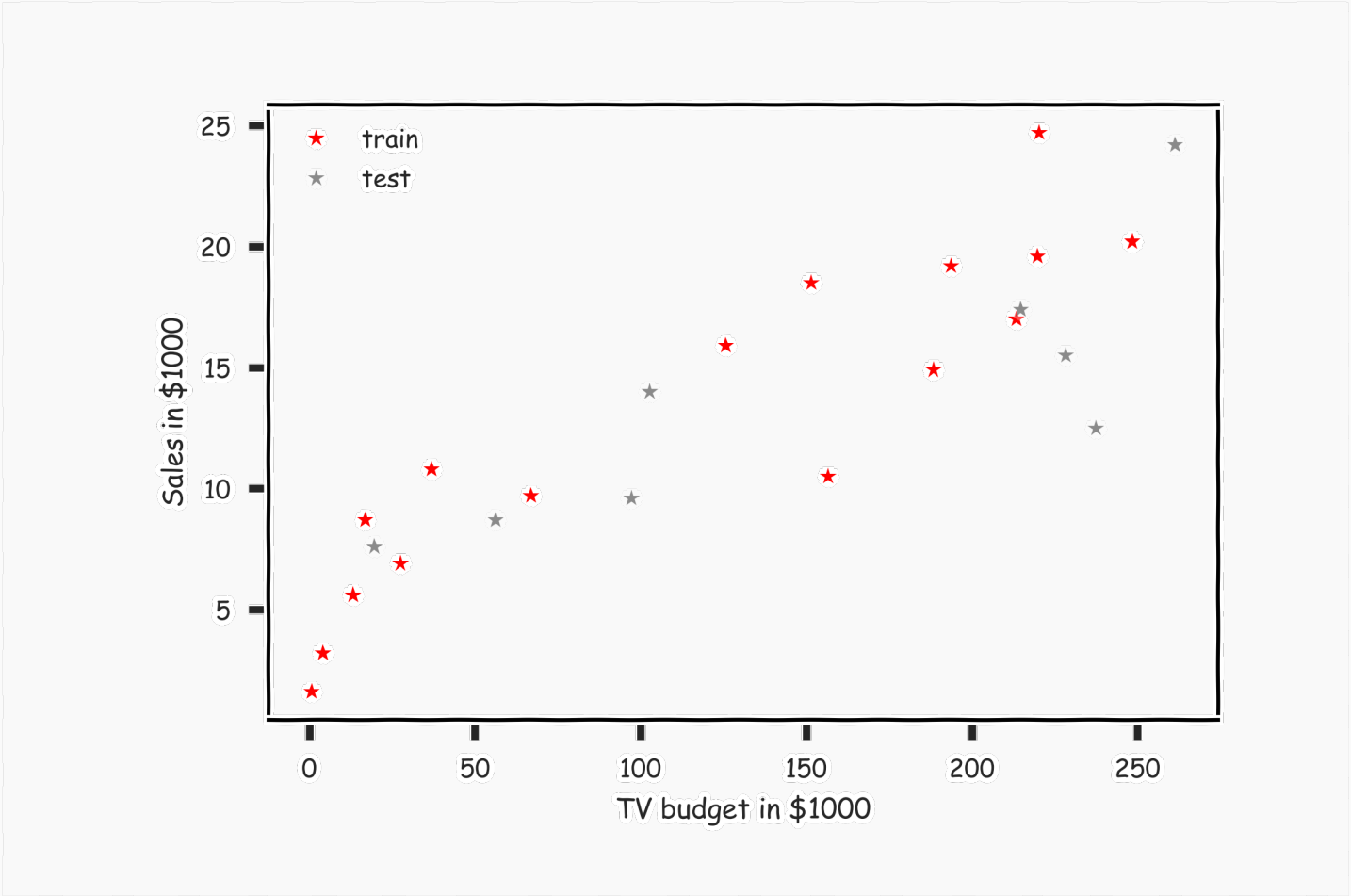
$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1^1 & \ldots & x_1^M \\ 1 & x_2^1 & \ldots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \ldots & x_n^M \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_M \end{pmatrix}.$$

Polynomial Regression

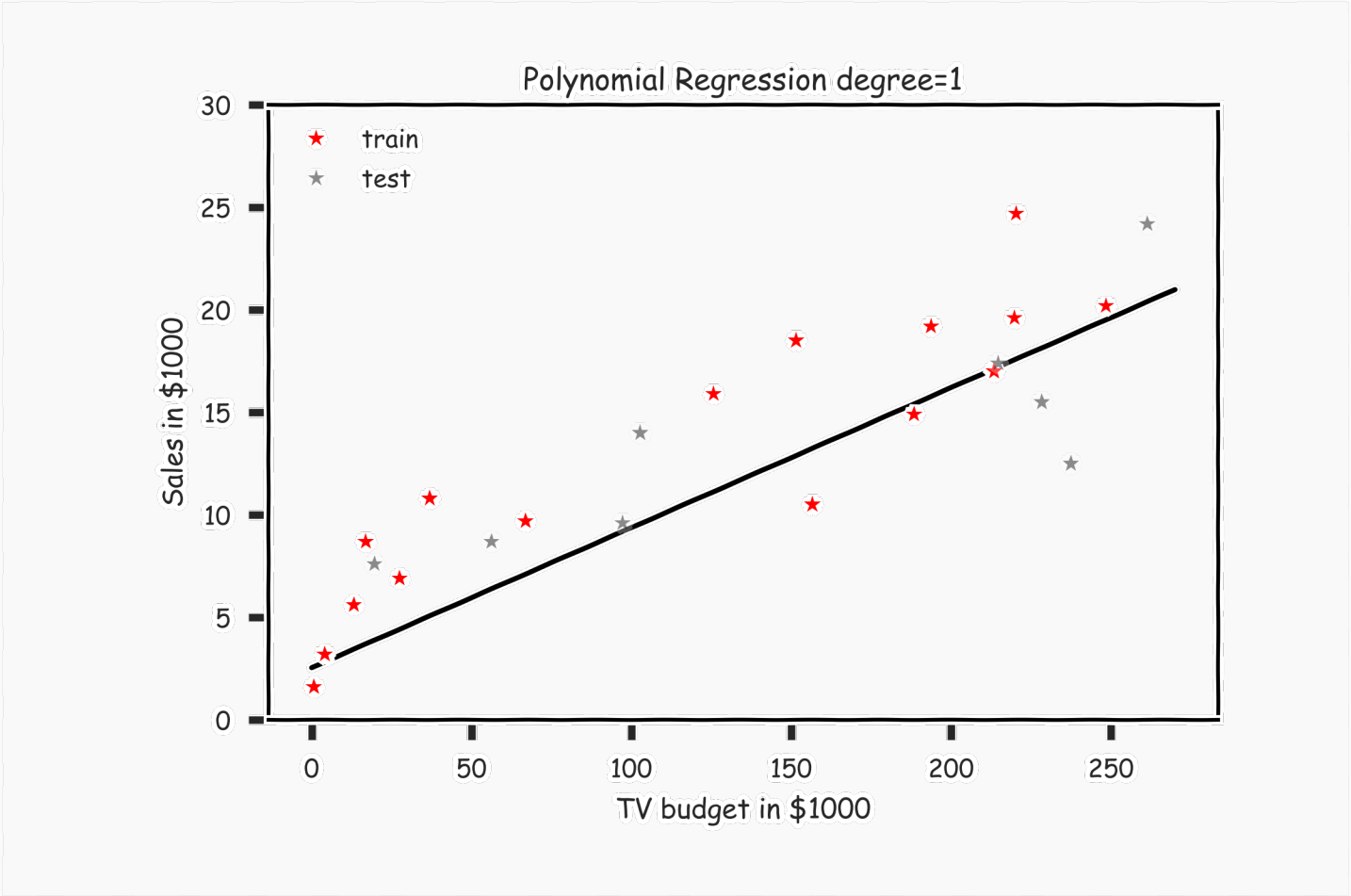
Again, minimizing the MSE using vector calculus yields,

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \operatorname{MSE}(\boldsymbol{\beta}) = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y}.$$

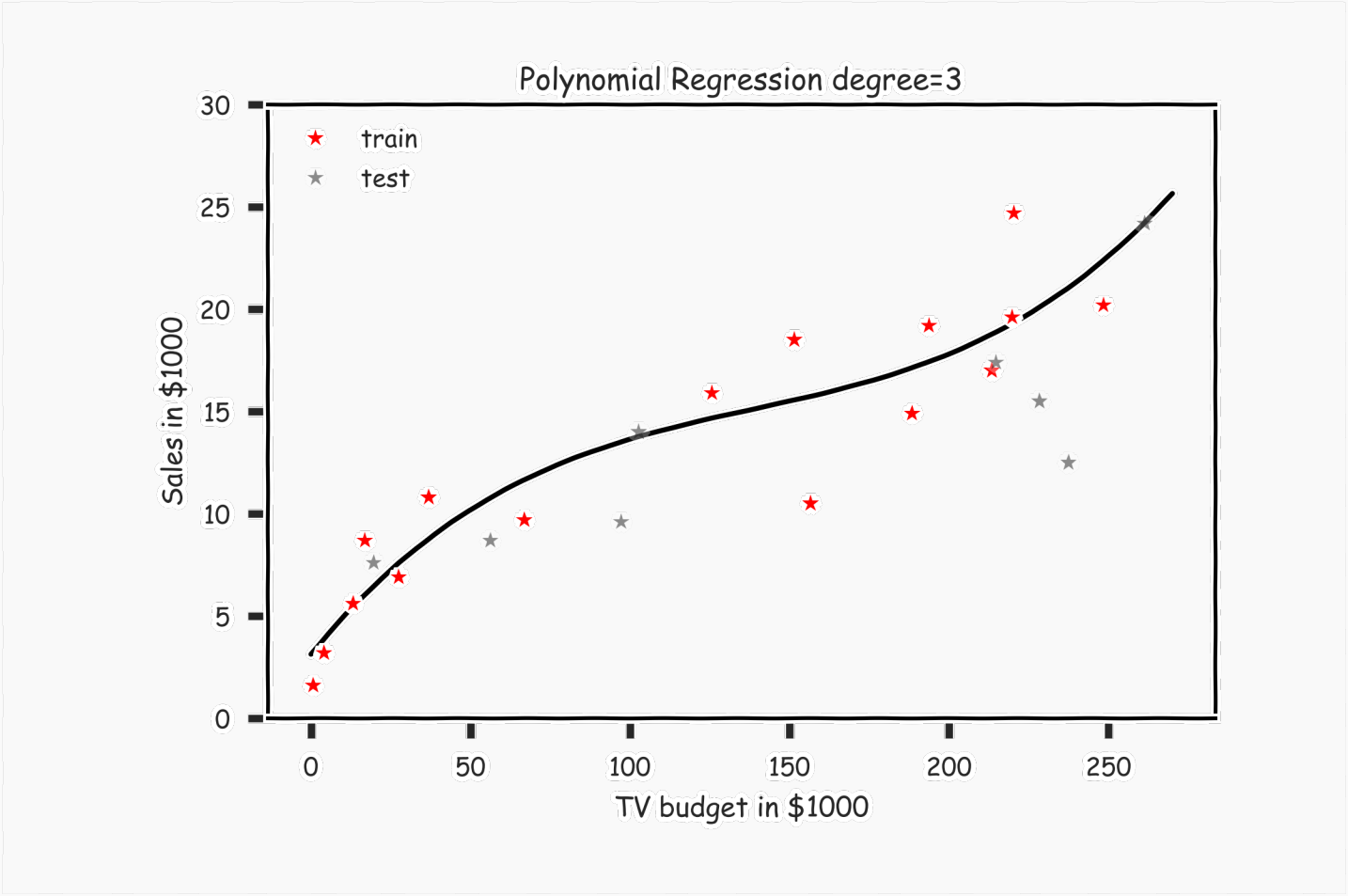
Polynomial Regression (cont)



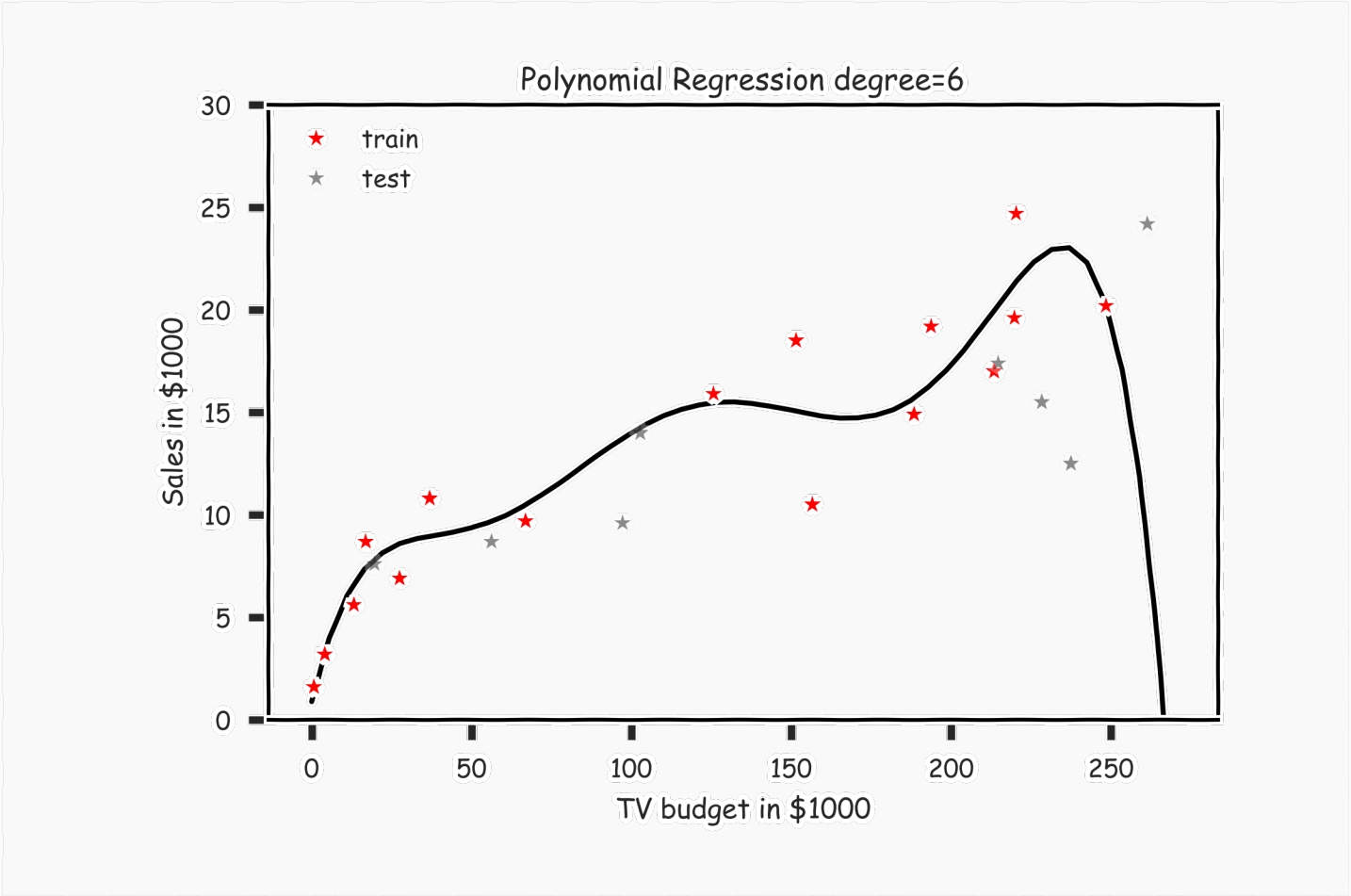
Polynomial Regression (cont)



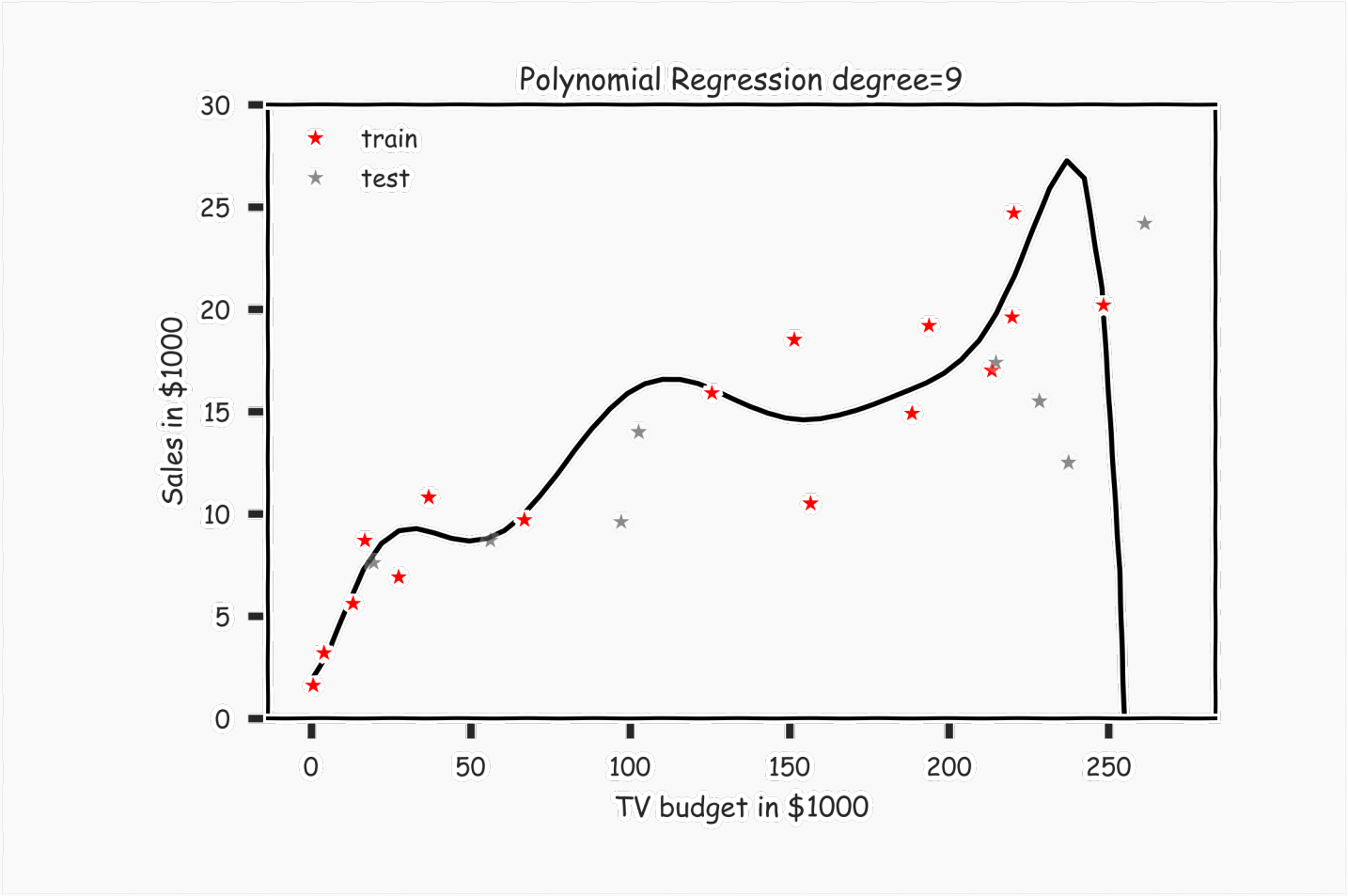
Polynomial Regression (cont)



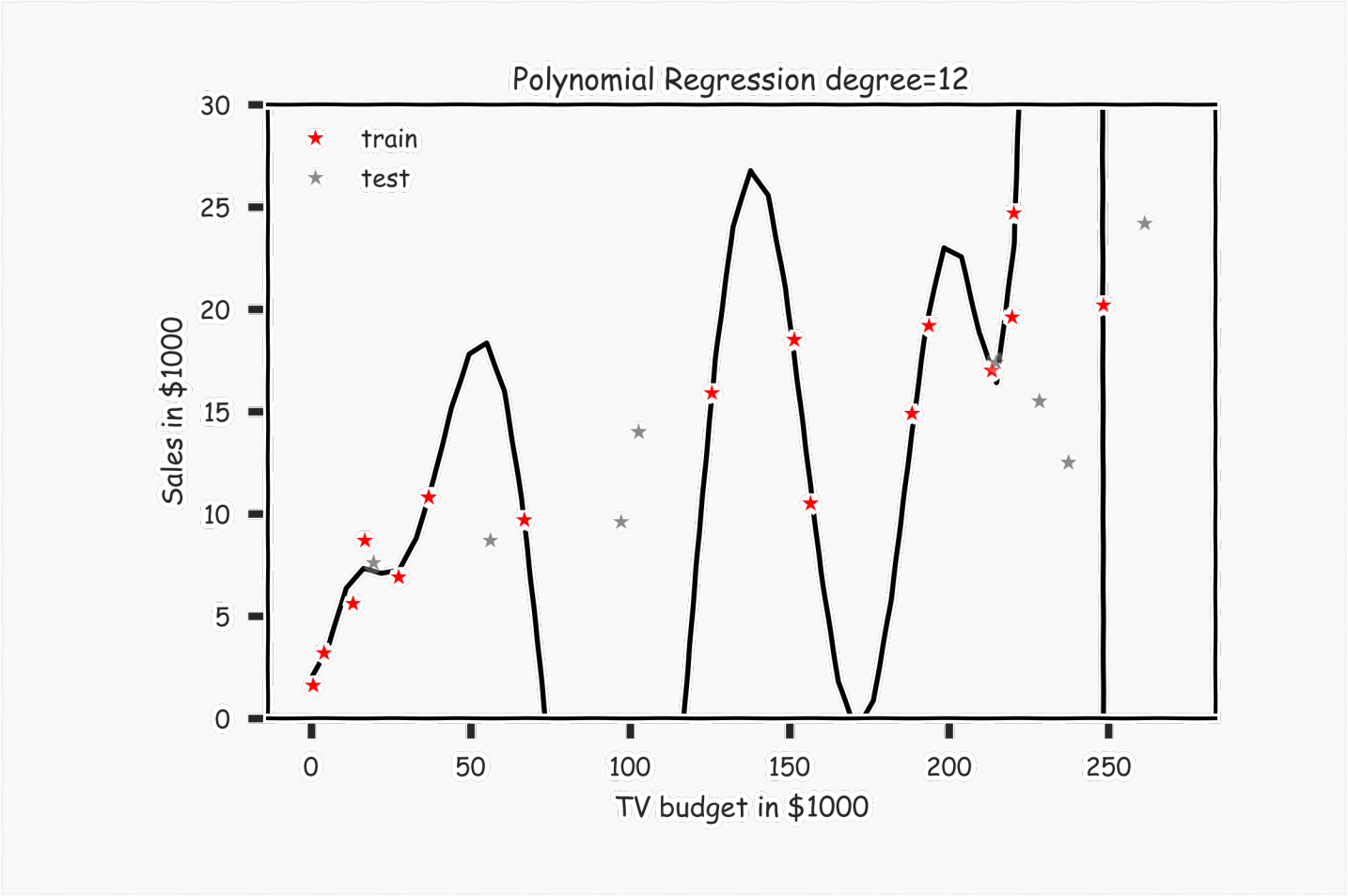
Polynomial Regression (cont)



Polynomial Regression (cont)



Polynomial Regression (cont)



Polynomial Regression

The simplest non-linear model we can consider, for a response *Y* and a predictor *X*, is a polynomial model of degree *M*,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_M x^M + \epsilon.$$

Just as in the case of linear regression with cross terms, polynomial regression is a special case of linear regression - we treat each x^m as a separate predictor. Thus, we can write

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1^1 & \ldots & x_1^M \\ 1 & x_2^1 & \ldots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \ldots & x_n^M \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_M \end{pmatrix}.$$

Polynomial Regression is Multiple Regression!

Multiple Regression

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,J} \\ 1 & x_{2,1} & \dots & x_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,J} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_J \end{pmatrix},$$

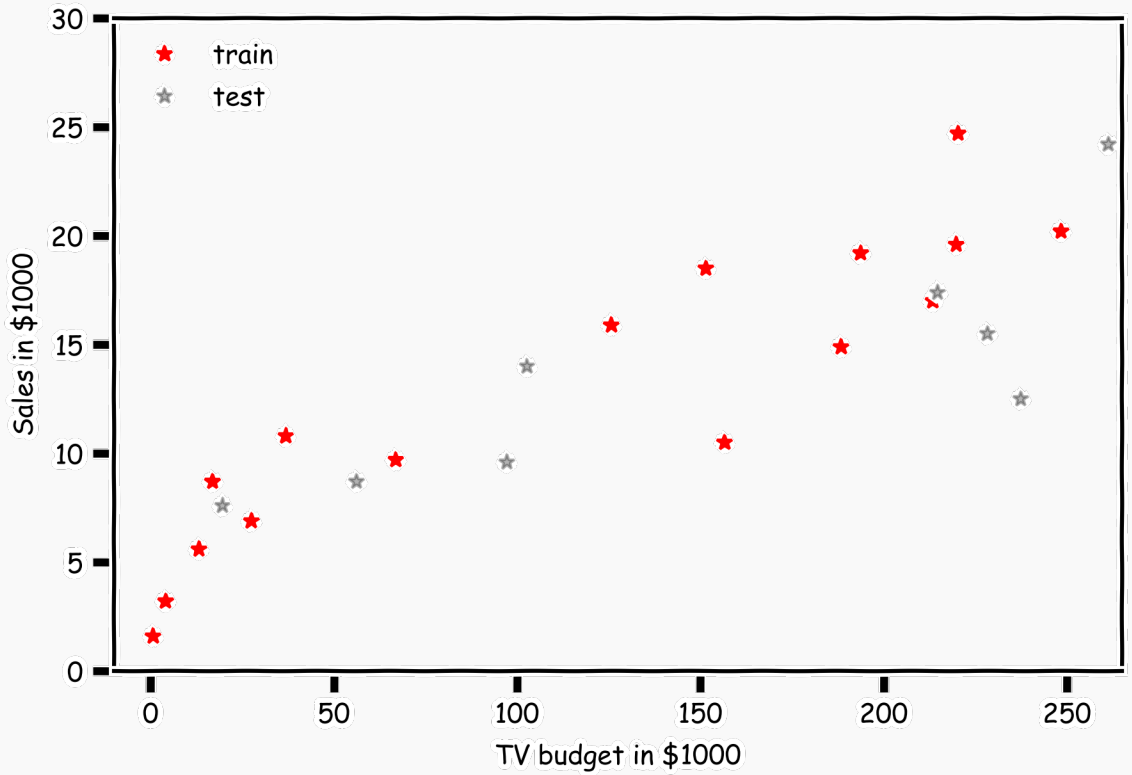
Poly-Regression

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1^1 & \dots & x_1^M \\ 1 & x_2^1 & \dots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \dots & x_n^M \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_M \end{pmatrix}.$$

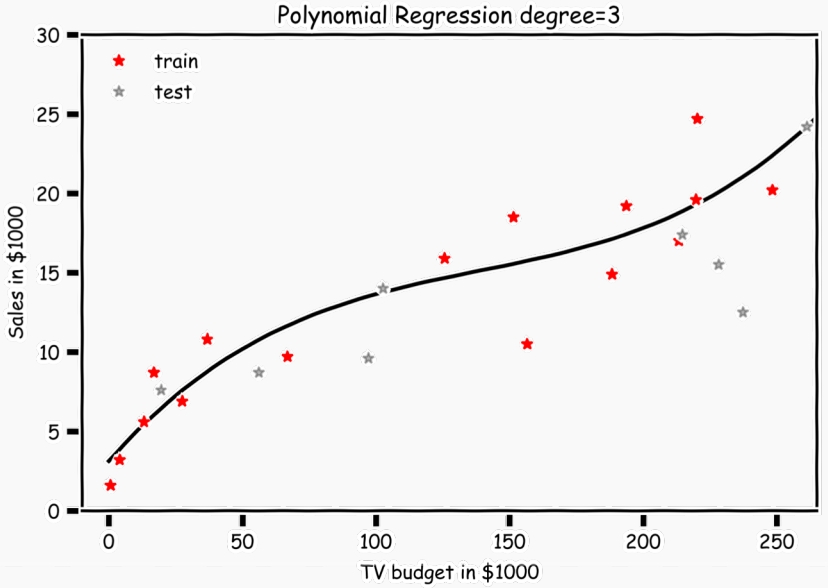
Lecture Outline

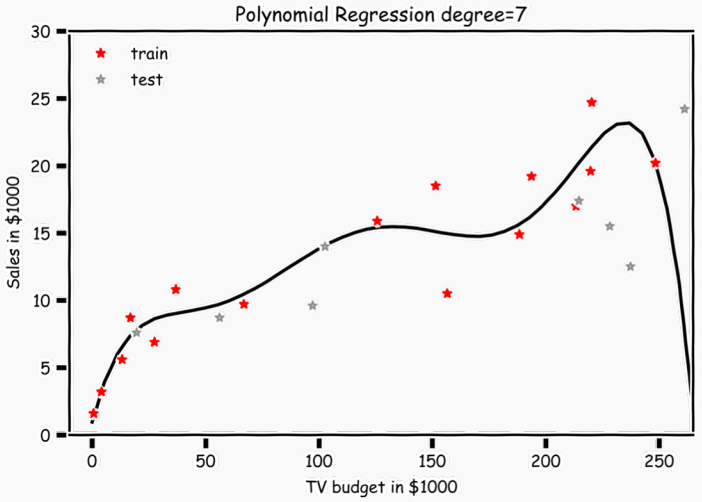
- Brief Recap
- How well do we know $\hat{f}$
  - Confidence and Prediction Intervals around our $\hat{f}$
- Multiple Regression
  - Formulation using Linear Algebra
- Categorical Predictors
- Interaction Terms
- Polynomial Regression
  - Linear Algebra Formulation
- **Overfitting**
- Variable Selection
- Bias vs. Variance

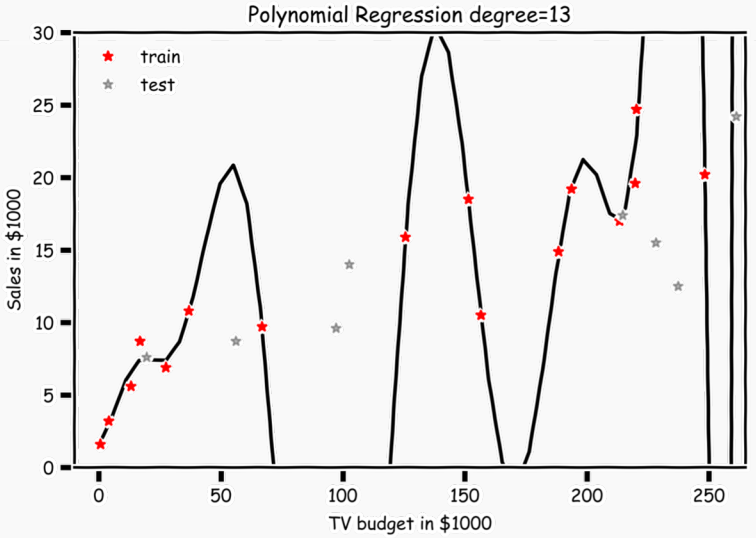
Overfitting



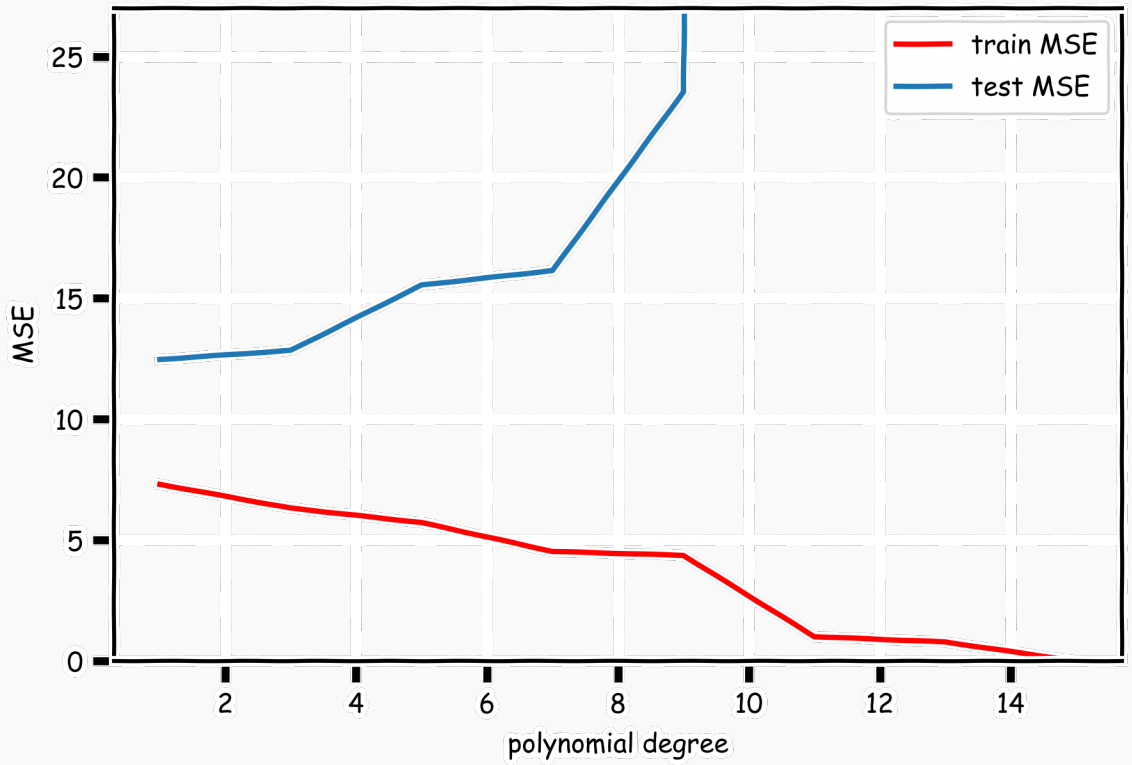
Overfitting with polynomials







Validation/Test



Overfitting

In statistical modeling, **overfitting** is *the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to properly capture the true relationships or predict future (aka, out-of-sample) observations reliably*.

You can think of overfitting as fitting to the particular intricacies of the residuals in the training set and not just the true signal in the relationship between the response and the predictors:

Think:

$$Y = f(\boldsymbol{X}) + \varepsilon$$

Train-Validation-Test

Question:

What is a more honest representation of the performance of the model (a polynomial model of degree = 10)?

R2_test = 0.52

R2_train = 0.83

Lecture Outline

- Brief Recap
- How well do we know $\hat{f}$
  - Confidence and Prediction Intervals around our $\hat{f}$
- Multiple Regression
  - Formulation using Linear Algebra
- Categorical Predictors
- Interaction Terms
- Polynomial Regression
  - Linear Algebra Formulation
- Overfitting
- **Variable Selection**
- Bias vs. Variance

Variable Selection

Variable selection (also called **model selection**) is the application of a principled method to determine the complexity of the model, e.g. choosing a subset of predictors, choosing the degree of the polynomial model etc.

A strong motivation for performing model selection is to avoid overfitting, which we saw can happen when:

- there are too many predictors:
    - the feature space has high dimensionality
    - the polynomial degree is too high
    - too many interaction terms are considered
- the coefficients values are too **extreme** (high multicollinearity)

Model Selection

Question:

How many different models could be used when considering a set of J predictors?

Model Selection

Example: J = 3 predictors (X_1, X_2, X_3)

- Models with 0 predictor:
  - M0: [intercept only]
- Models with 1 predictor:
  - M1: X_1
  - M2: X_2
  - M3: X_3
- Models with 2 predictors:
  - M4: $\{X_1, X_2\}$
  - M5: $\{X_2, X_3\}$
  - M6: $\{X_3, X_1\}$
- Models with 3 predictors:
  - M7: $\{X_1, X_2, X_3\}$

$2^J = 8$ Models

Stepwise Variable Selection (vs. Cross-Validation)

Selecting optimal subsets of predictors (including choosing the degree of polynomial models) through:

- **stepwise variable selection** - iteratively building an optimal subset of predictors by optimizing a fixed model evaluation metric each time,
- **Cross-validation** - selecting an optimal model by evaluating each model on validation/test set(s)…more on this later.

We will also address the issue of discouraging extreme values in model parameters later.

Stepwise Variable Selection: Forward method

In **forward selection**, we find an 'optimal' set of predictors by iterative building up our set.

1. Start with the empty set P0, construct the null modelM0.

2. For $k = 1, \ldots, J$:

2.1 Let M_{k-1} be the model constructed from the best set of $k-1$ predictors, P_{k-1}.

2.2 Select the predictor X_{n_k}, not in P_{k-1}, so that the model constructed from $P_k = X_{n_k} \cup P_{k-1}$ optimizes a fixed metric (this can be *p*-value, *F*-stat; validation/test MSE, R^2; or AIC/BIC on training set).

2.3 Let M_k denote the model constructed from the optimal P_k.

3. Select the model M amongst $\{M_0, M_1, \ldots, M_J\}$ that optimizes a fixed metric (this can be validation MSE, R^2; or AIC/BIS on training set)

*Note: this reverse direction (starting with the *full* model) is also possible: *Backwards* Stepwise

Stepwise Variable Selection Computational Complexity

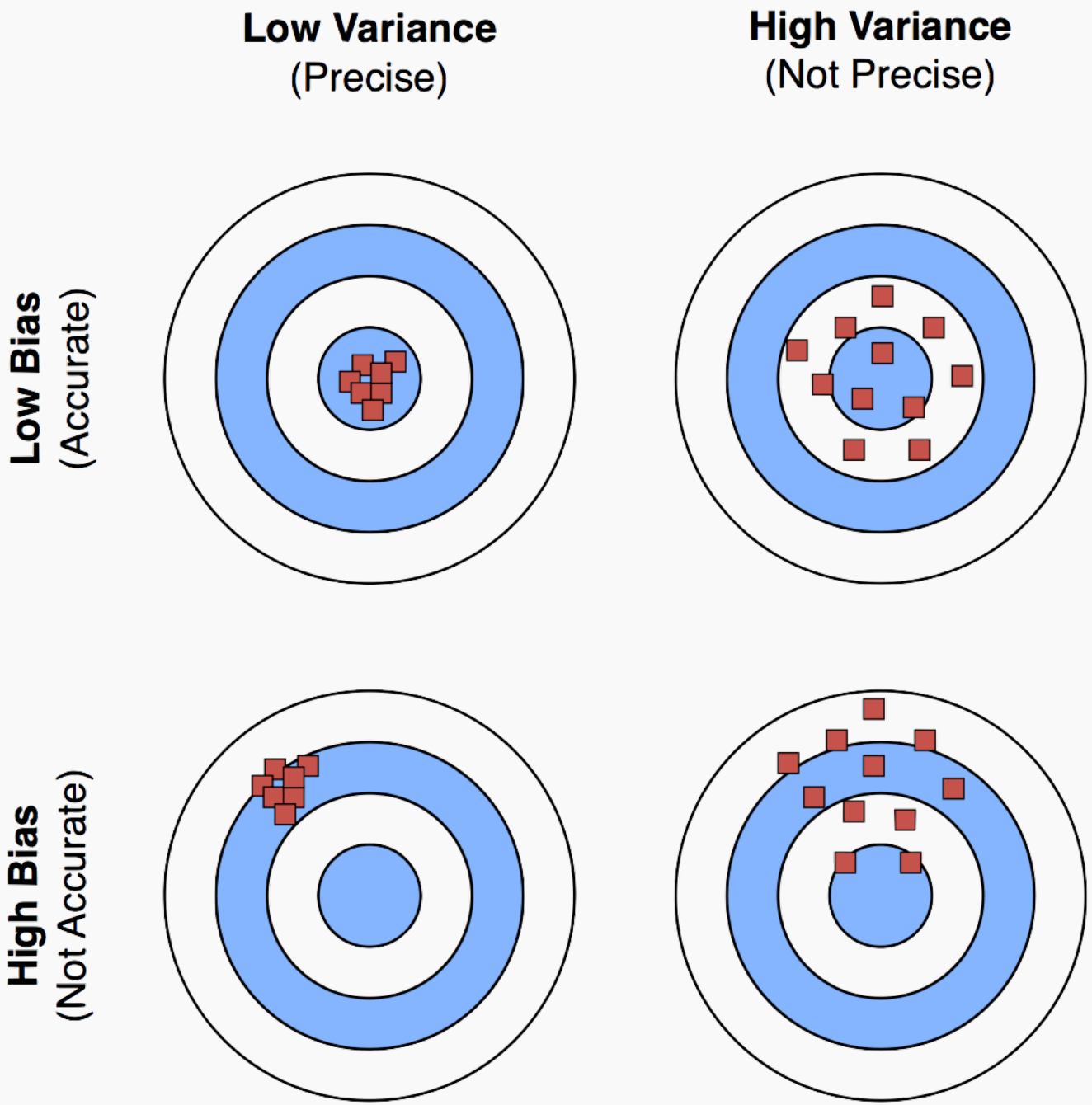
How many models did we evaluate?

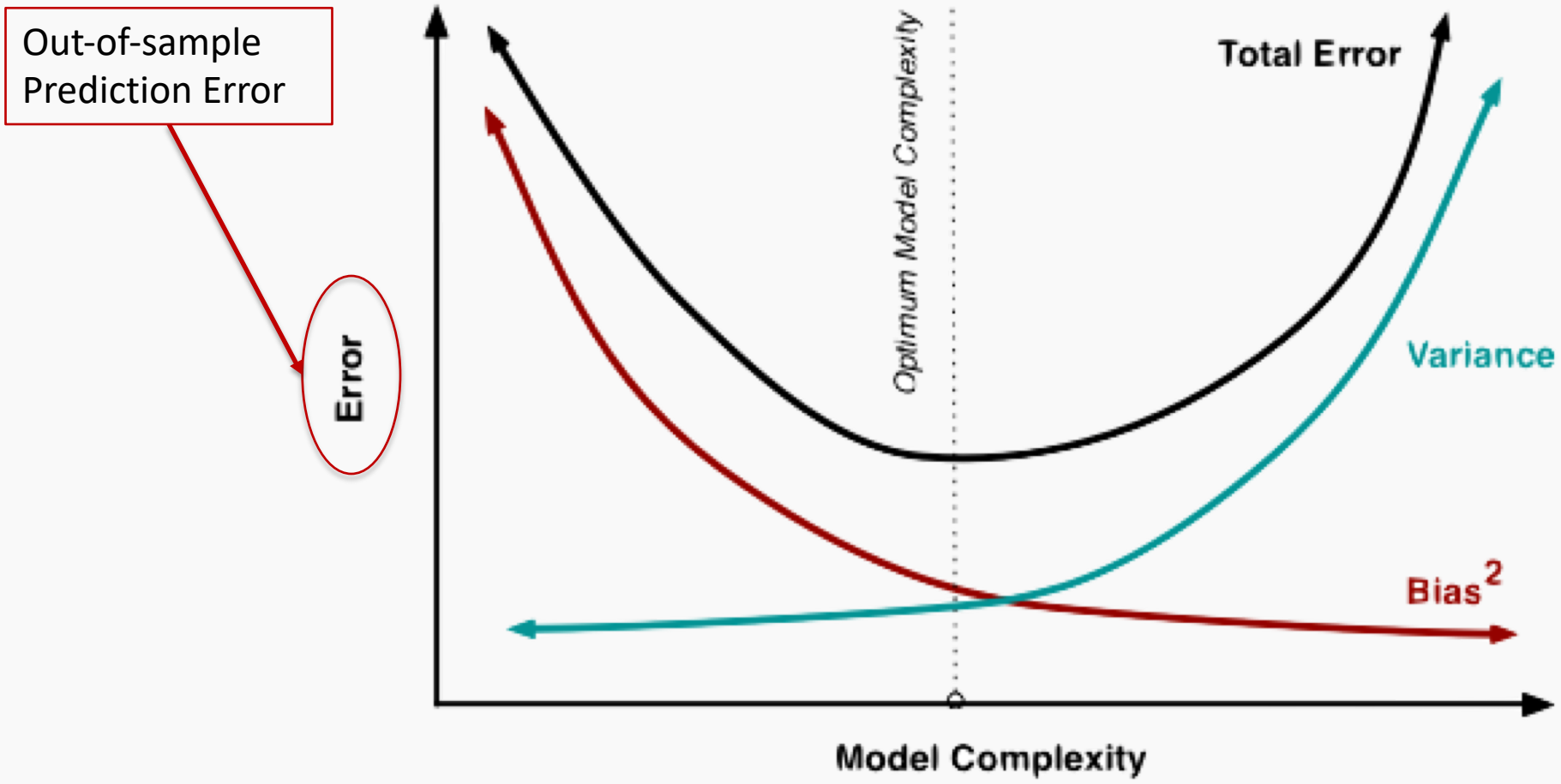
- 1st step, ***J* Models**
- 2nd step, *J-1* **Models** (add 1 predictor out of *J-1* possible)
- 3rd step, *J-2* **Models** (add 1 predictor out of *J-2* possible)
- …

$$O(J^2) \ll 2^J \text{ for large } J$$

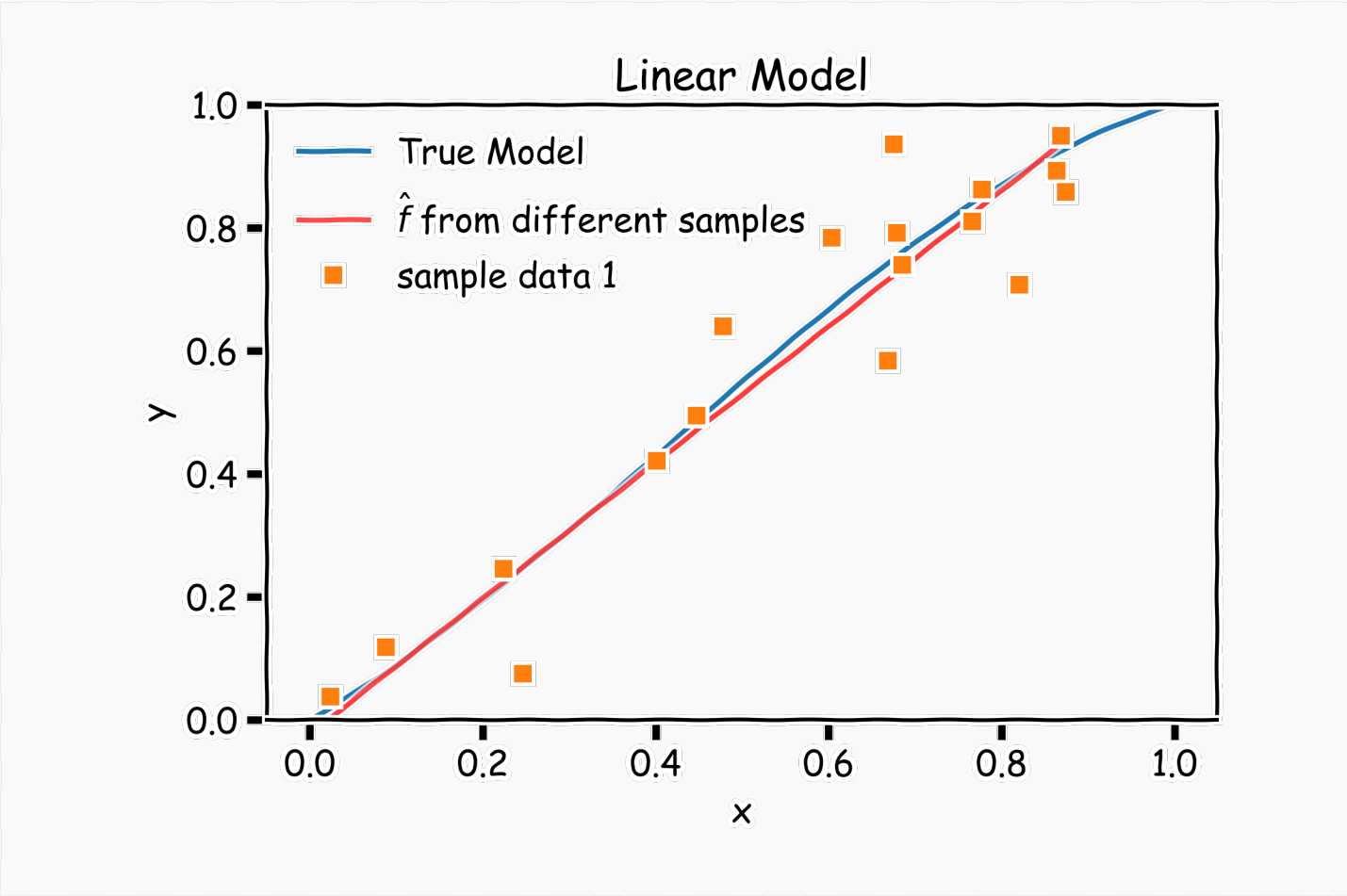
Lecture Outline

- Brief Recap
- How well do we know $\hat{f}$
  - Confidence and Prediction Intervals around our $\hat{f}$
- Multiple Regression
  - Formulation using Linear Algebra
- Categorical Predictors
- Interaction Terms
- Polynomial Regression
  - Linear Algebra Formulation
- Overfitting
- Variable Selection
- **Bias vs. Variance**

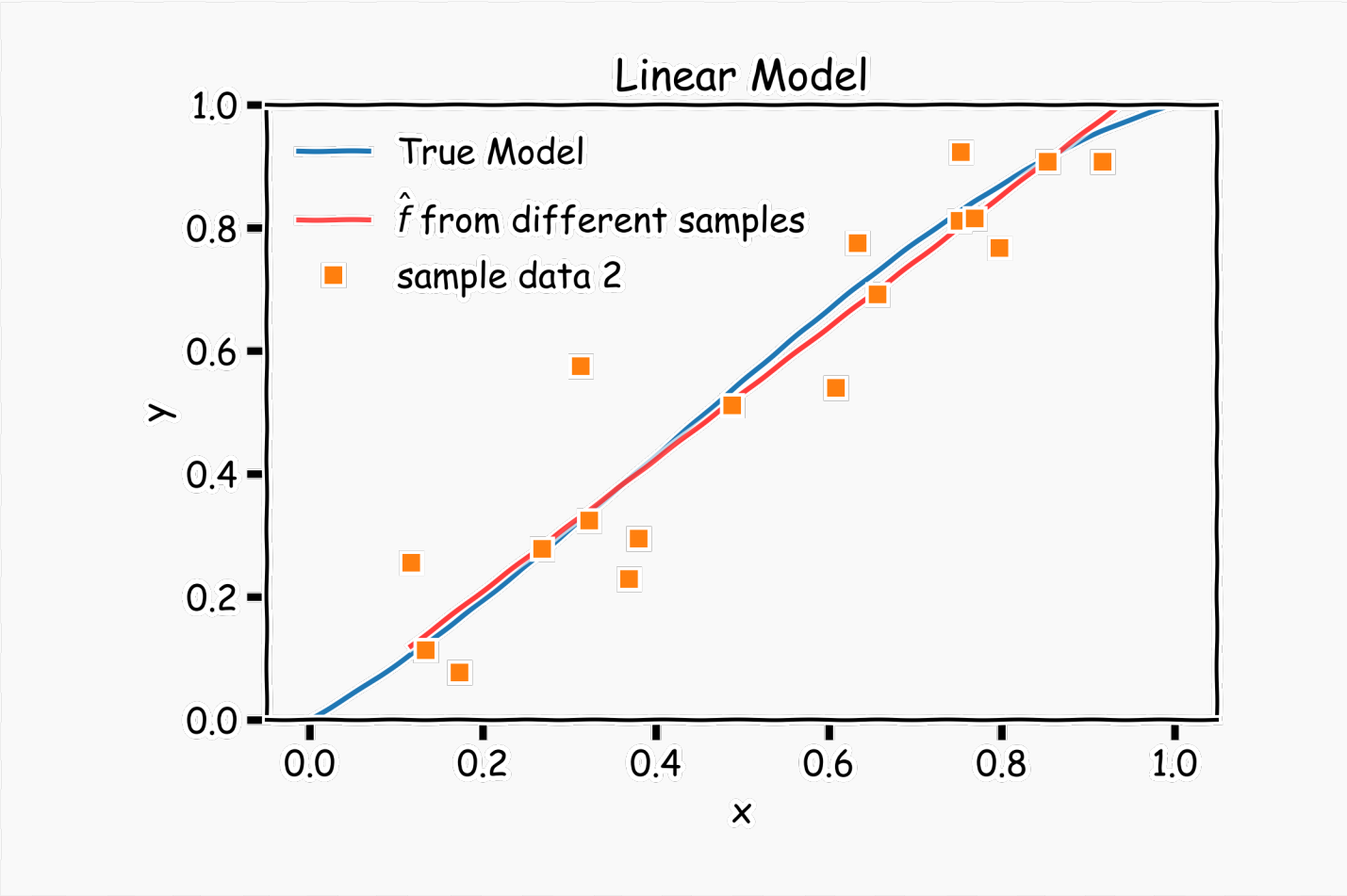

Low Variance
(Precise)
High Variance
(Not Precise)
Low Bias
(Accurate)
High Bias
(Not Accurate)


Out-of-sample
Prediction Error
Error
Optimum Model Complexity
Total Error
Variance
Bias^2
Model Complexity

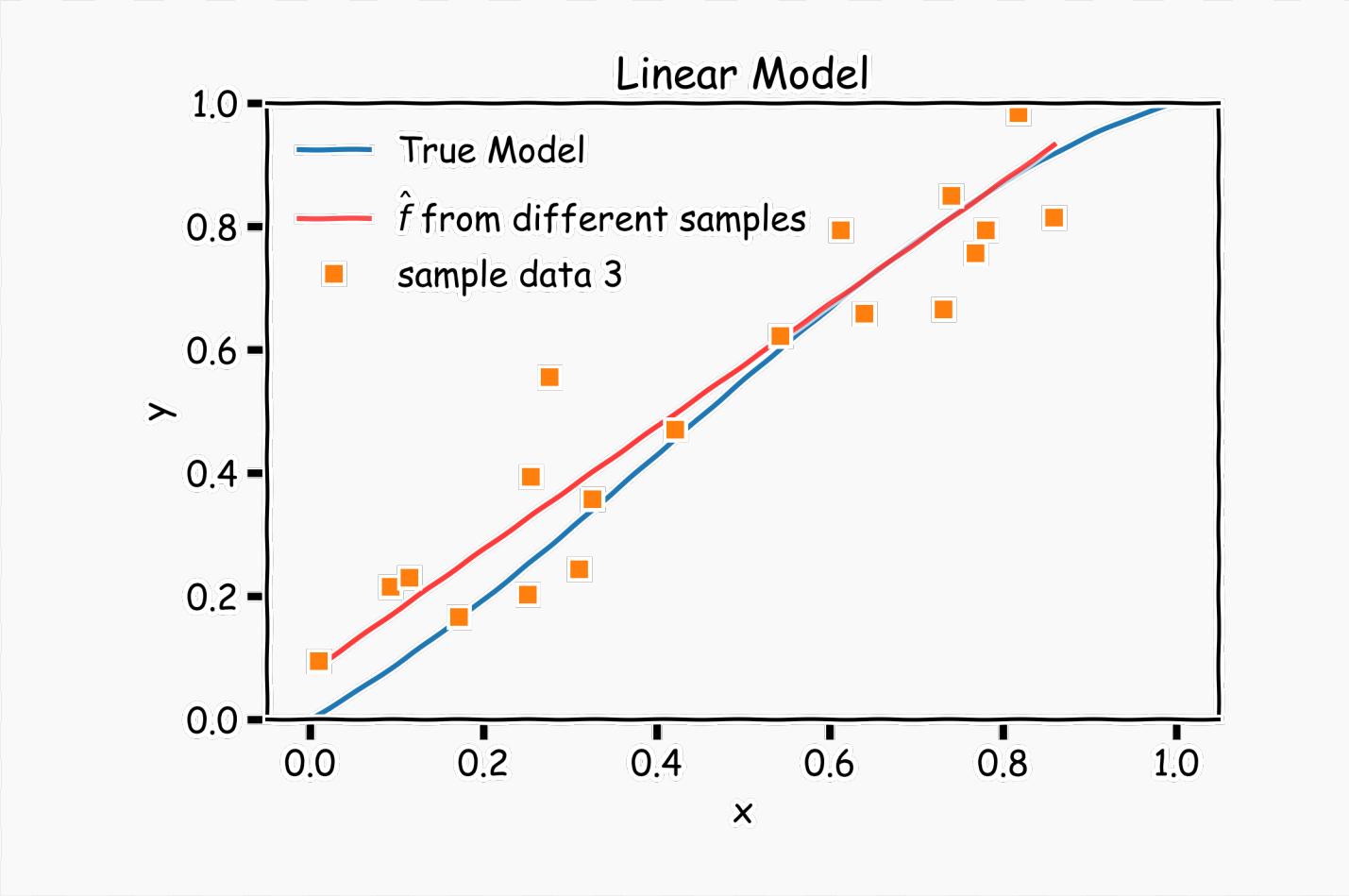
Bias vs Variance



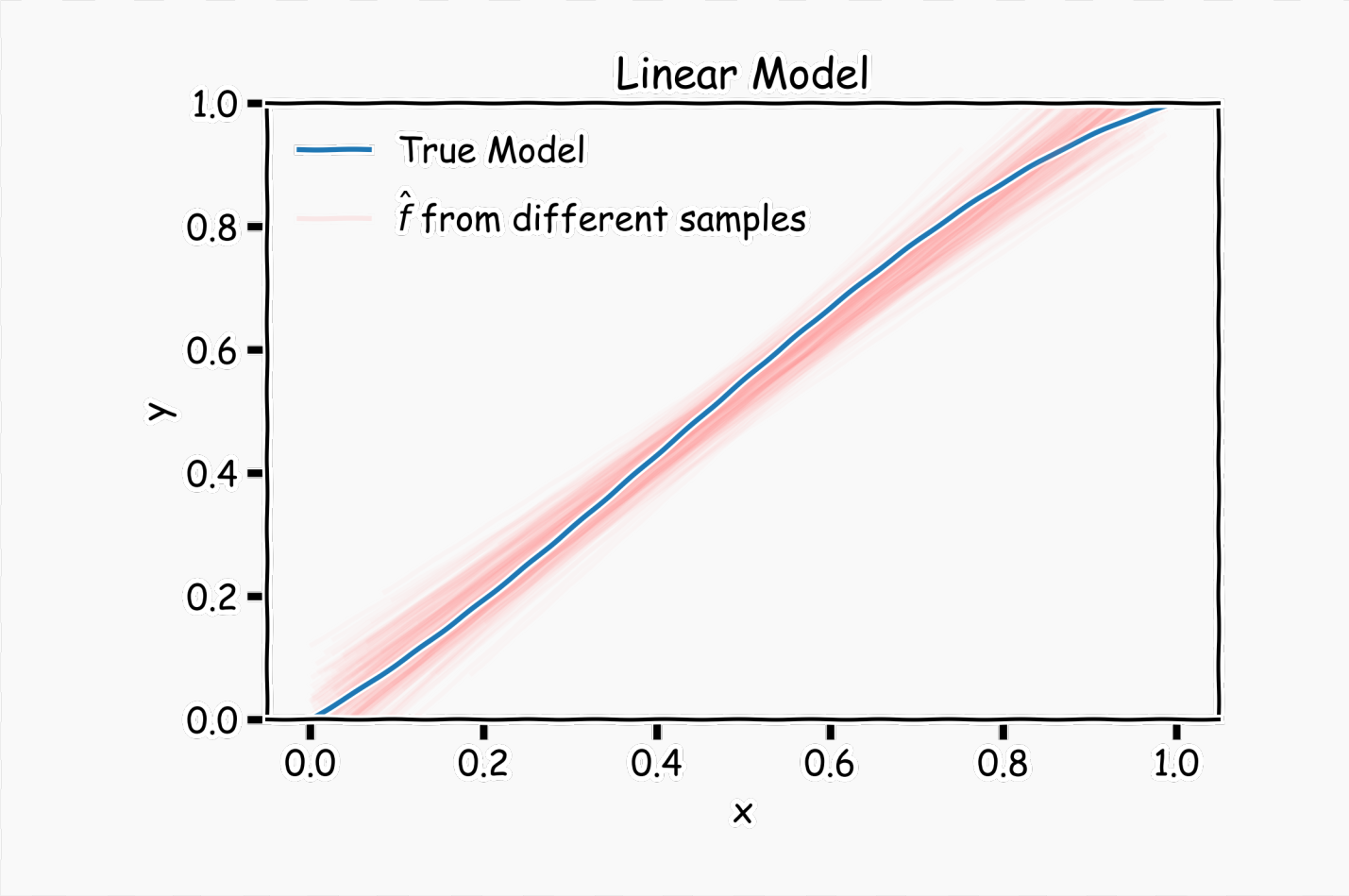
Bias vs Variance



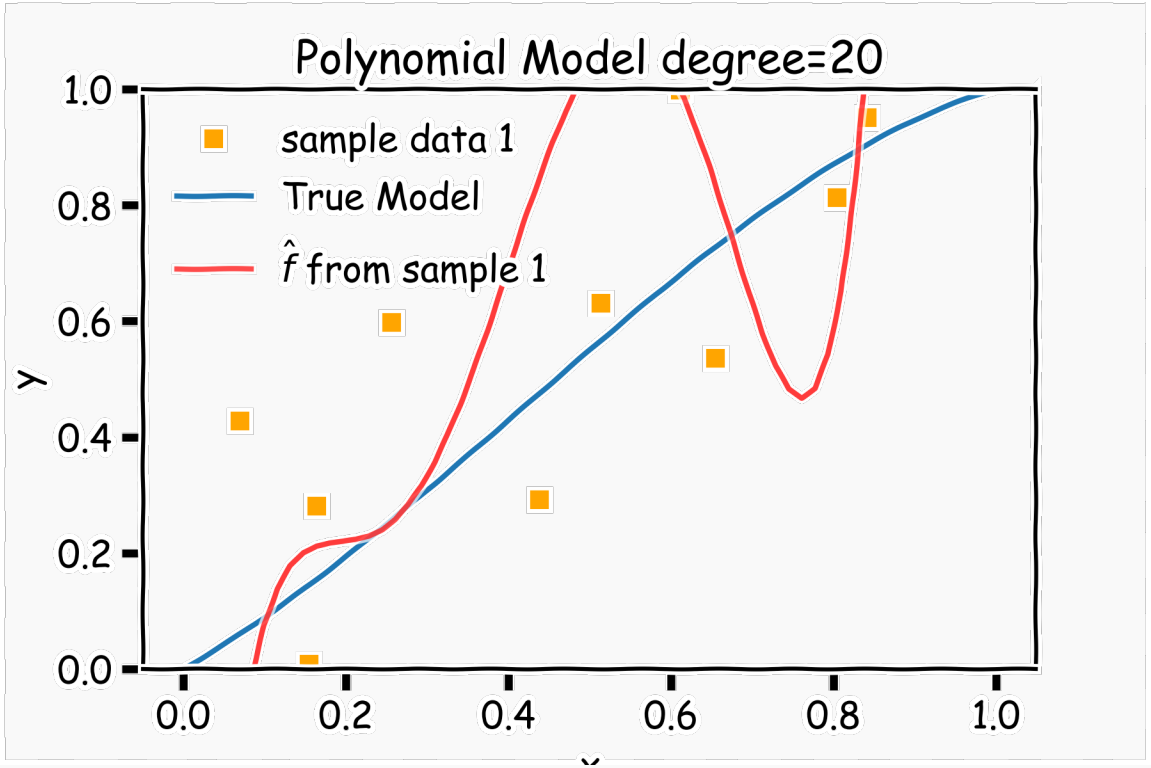
Bias vs Variance



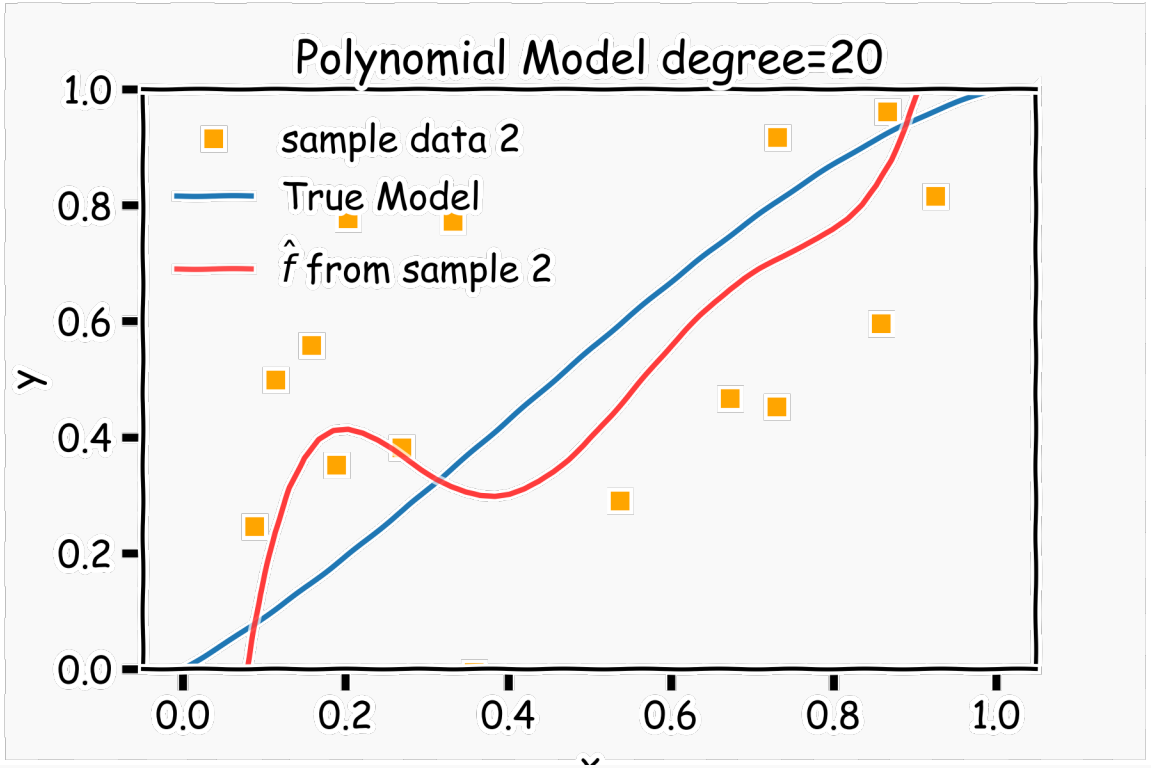
Linear models: 20 data points per line 2000 simulations



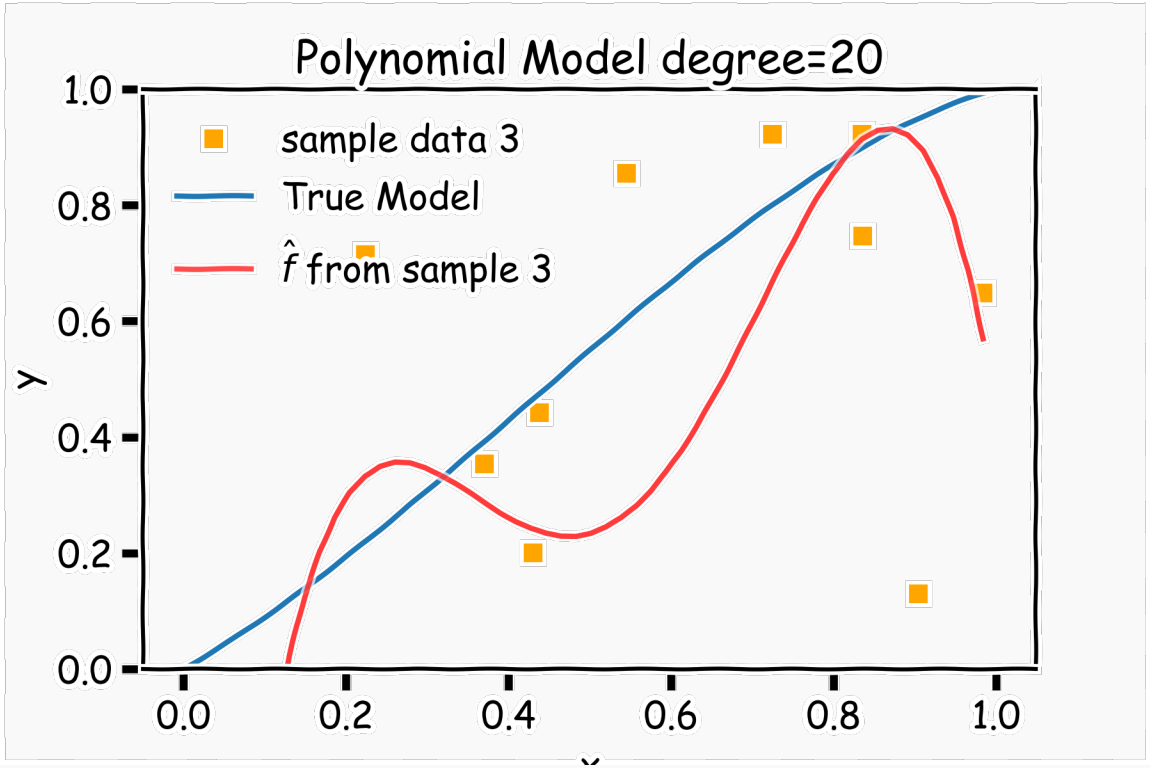
Bias vs Variance



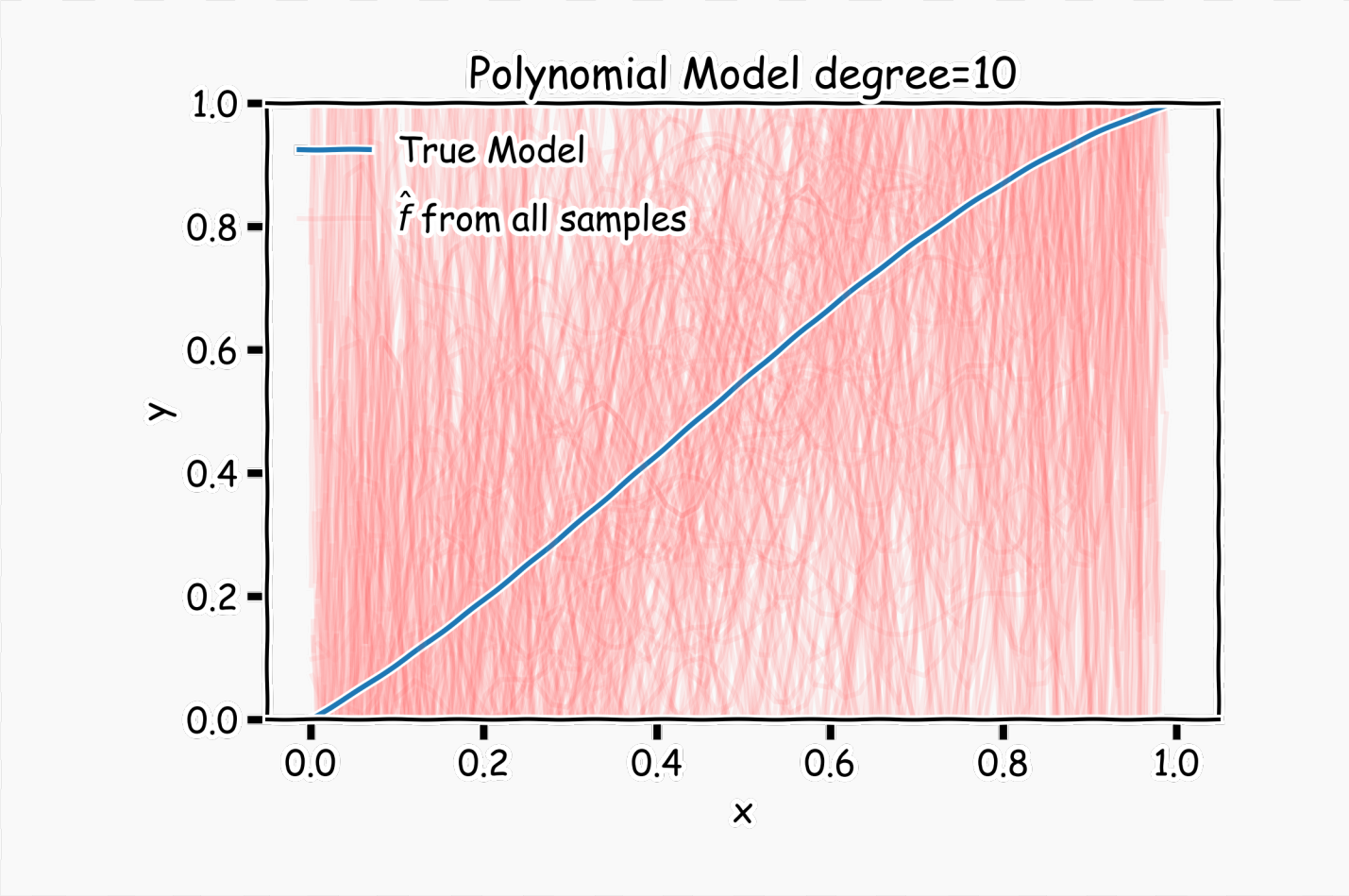
Bias vs Variance



Bias vs Variance



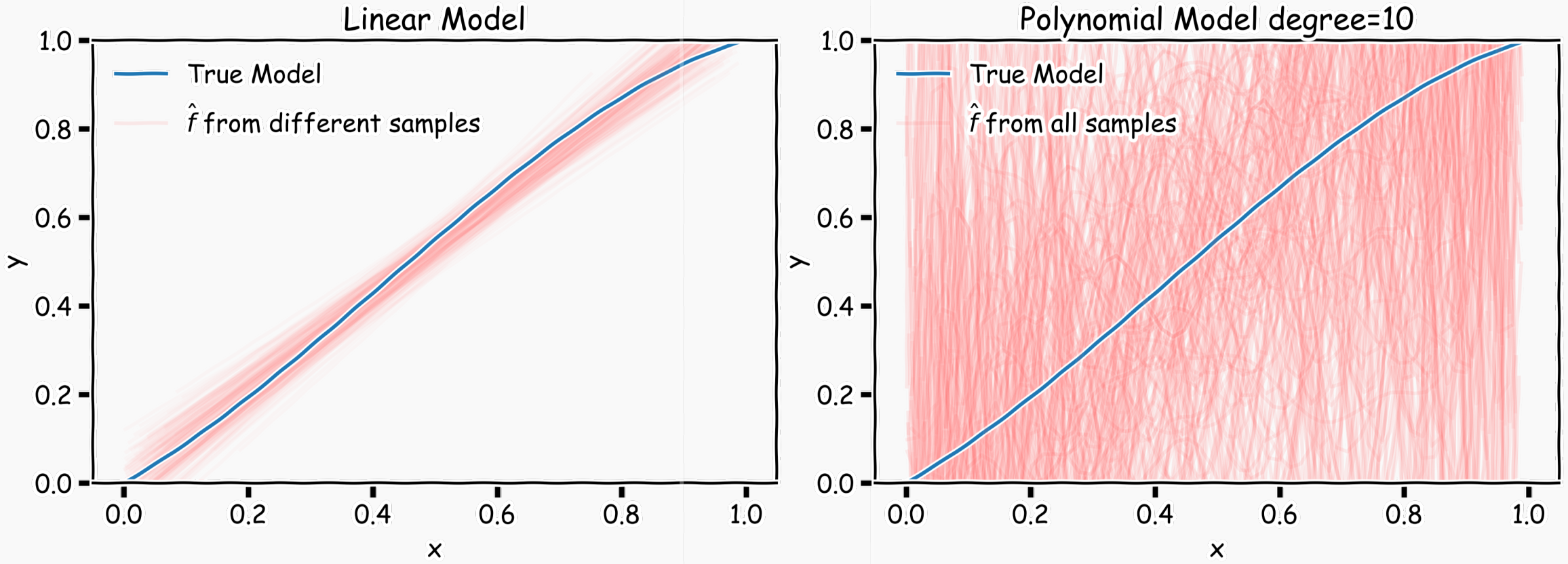
Poly 10 degree models : 20 data points per line 2000 simulations



Bias vs Variance

Left: 2000 best fit straight lines, each fitted on a different 20 point training set.

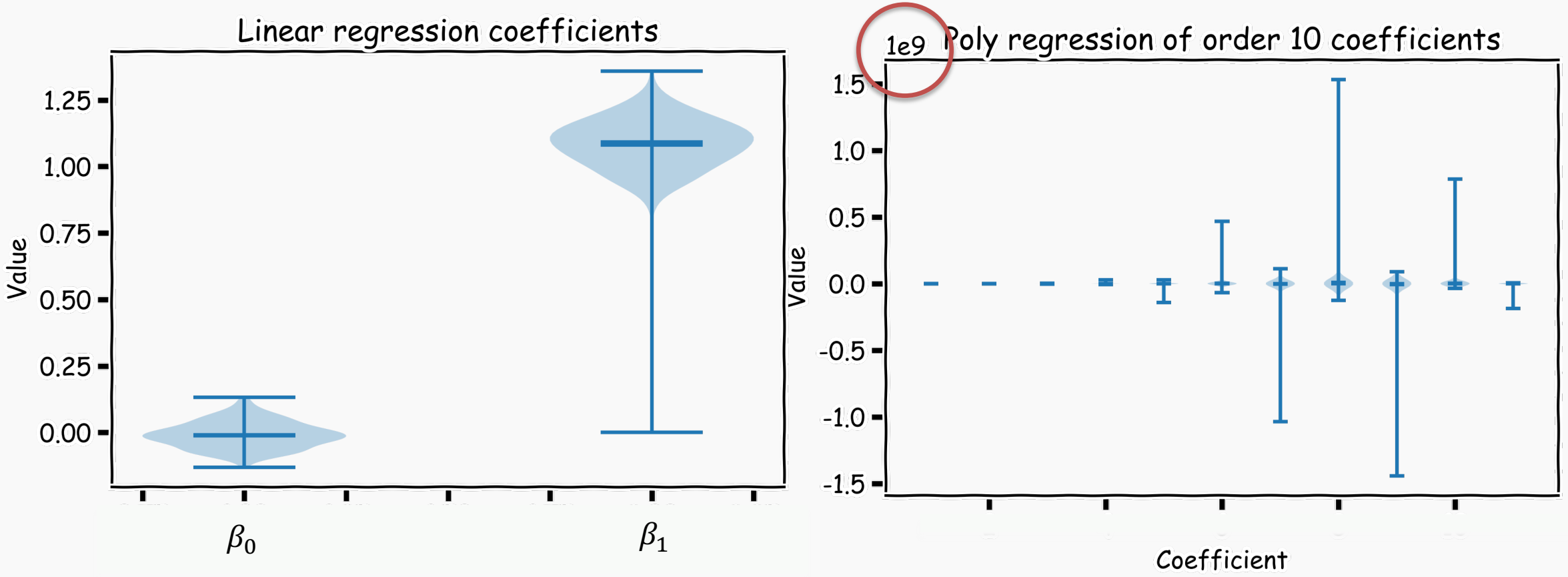
Right: Best-fit models using degree 10 polynomial



Bias vs Variance

Left: Linear regression coefficients

Right: Poly regression of order 10 coefficients



Bias vs Variance

Left: Linear regression coefficients

Right: Poly regression of order 10 coefficients