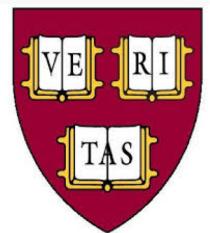
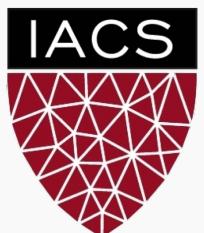


multiple predictors

## Lecture #4: Multiple Linear Regression

CS-S109A: Introduction to Data Science

Kevin Rader



HARVARD  
Summer School

## ANNOUNCEMENTS

- **HW2** is available, due Tues, July 7 at 11:59pm
- **Lab2** is tomorrow 12-2pm, not Friday due to the holiday.
- **Lec3 Quiz (the feedback survey)** will be left available until tomorrow morning. Please fill out for ‘free’ full credit quiz. (Lec4 Quiz will be released after class, available for 36+ hours).

Look for  
an  
announcement



# Lecture Outline

- Continuation  
of  
Simple  
regression*
- Bootstrap Resampling
  - 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



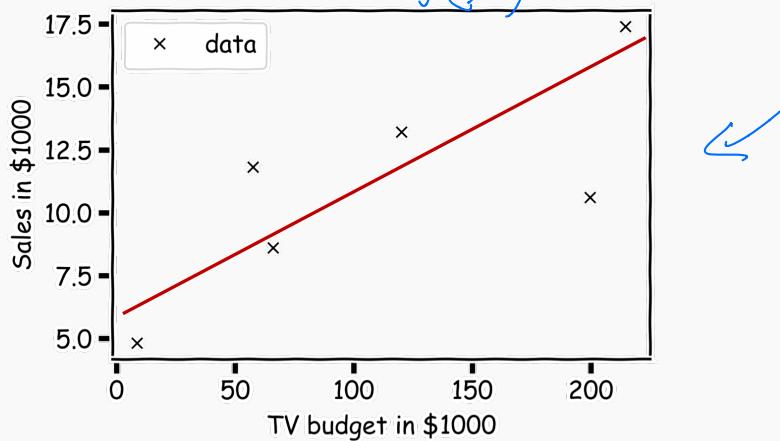
## Summary from last lecture

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

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

signal noise

$f(x)$



## Summary from last lecture

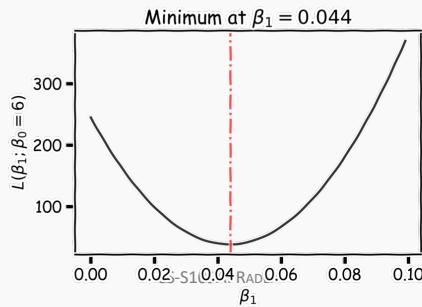
estimated coefficients:  
intercept and slope

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:

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

$\hookrightarrow$  squared deviations!  
 $(y_i - \hat{y}_i)^2$

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



## 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  $\hat{\beta}_0, \hat{\beta}_1$

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

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

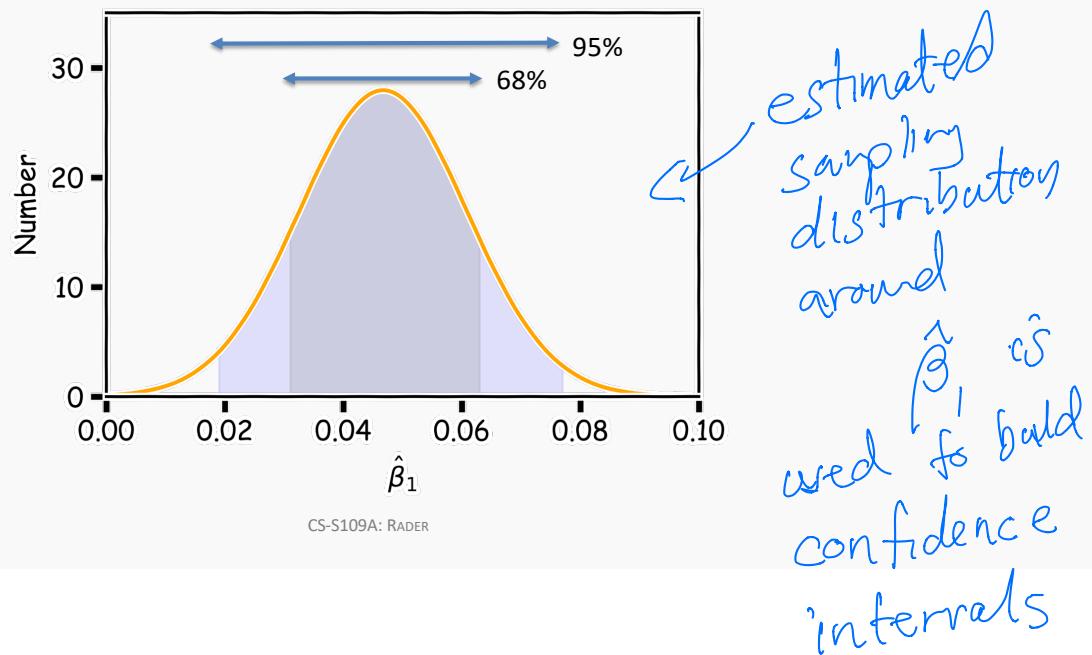
Sampling distributions  
of our estimates

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



## Summary from last lecture

We calculate the confidence intervals, which are the ranges of values such that the **true** value of  $\beta_1$  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.  $H_A: \beta_1 \neq 0$  (in simple regression):

$$T = \frac{\beta_1 - 0}{\sqrt{\hat{\sigma}^2 / \sum(x_i - \bar{x})^2}}$$

$\leftarrow t\text{-test}$

*no relationship*      *is a relationship*

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?



$$\begin{aligned} &\text{instead of } H_0: \beta_1 = 0 \\ &H_0: \beta_1 = 1 \end{aligned}$$

## 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*).

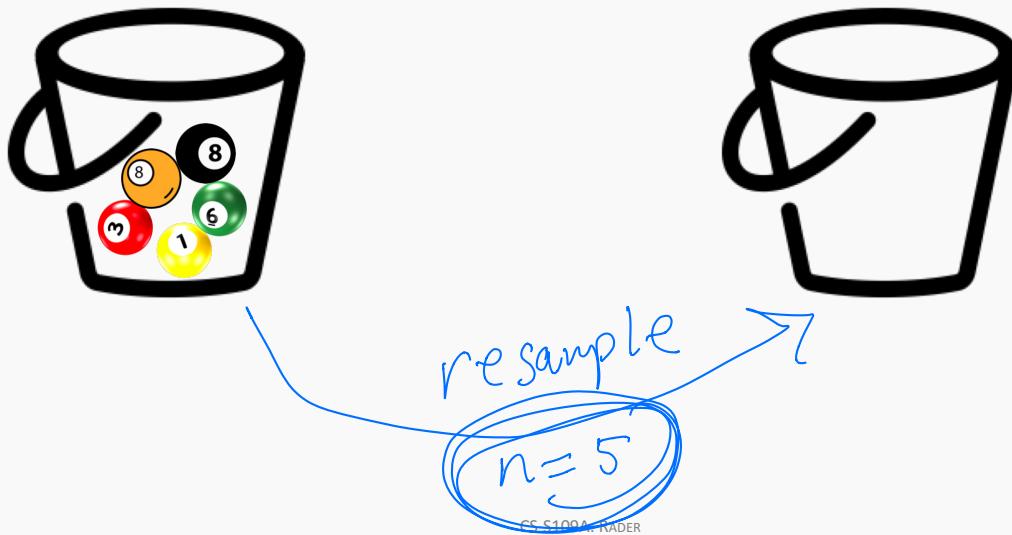
- estimate standard errors ( $\text{var}(\hat{\beta}_1)$ )
- build confidence intervals  
(without having to rely  
on normality assumptions)



## Bootstrap

represent an observation  
(a row)  
in our sample  
( $n=5$ )

Imagine we have 5 billiard balls in a bucket.



# Bootstrap

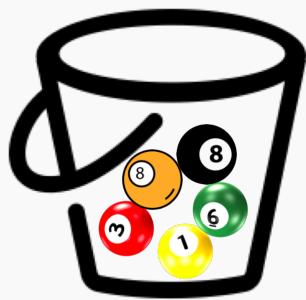
---

We first pick randomly a ball and replicate it. This is called **sampling with replacement**. We move the replicated ball to another bucket.



# Bootstrap

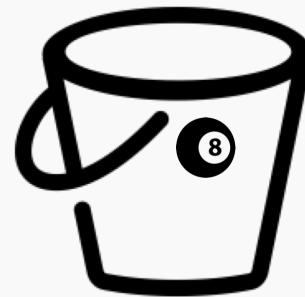
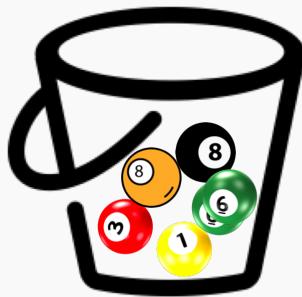
---



# Bootstrap

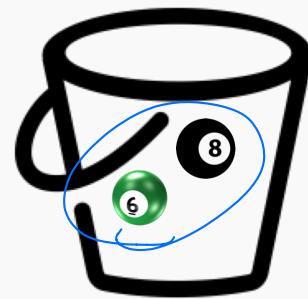
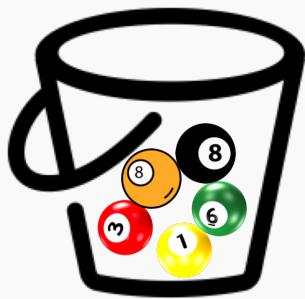
---

We then randomly pick another ball and again we replicate it.  
As before, we move the replicated ball to the other bucket.



# Bootstrap

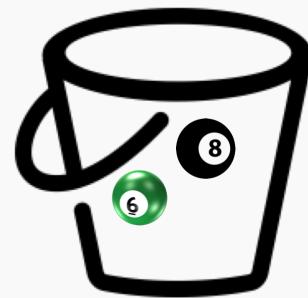
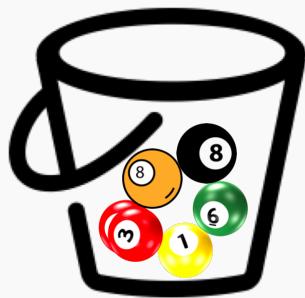
---



# Bootstrap

---

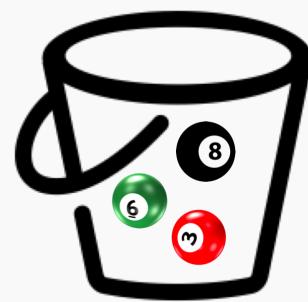
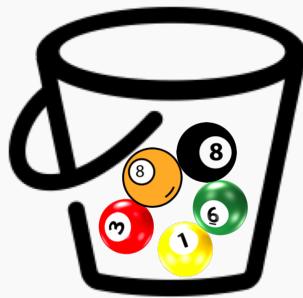
We repeat this process.



# Bootstrap

---

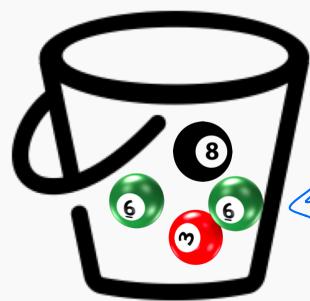
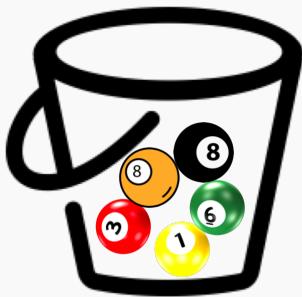
Again



# Bootstrap

---

And again

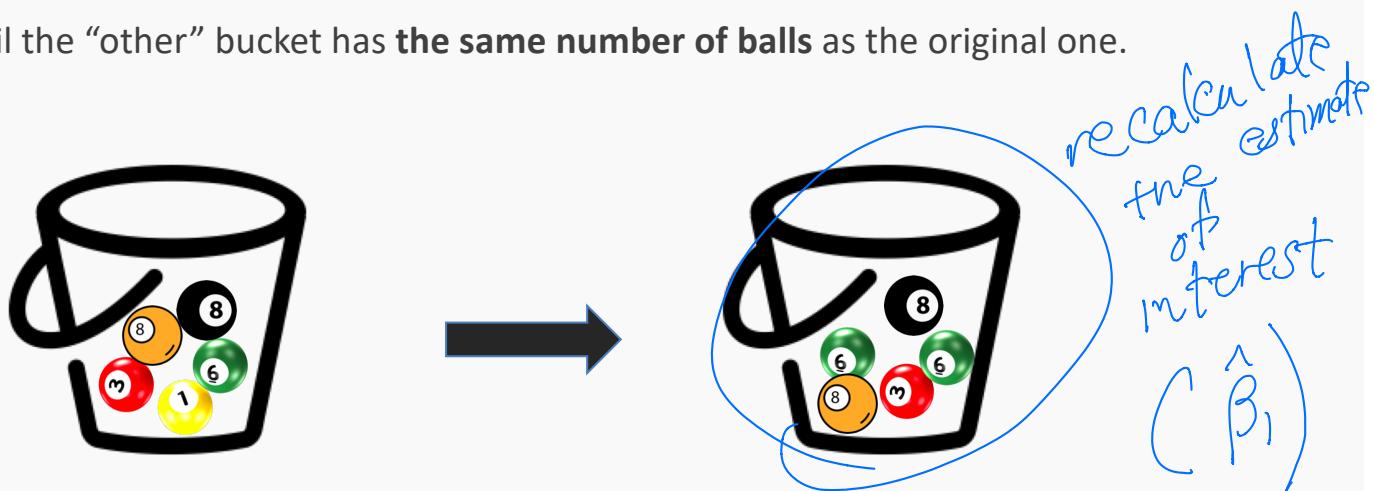


there  
are  
duplicates  
(and  
that's  
ok)



## Bootstrap

Until the “other” bucket has **the same number of balls** as the original one.



This new bucket represents a new parallel universe



# Bootstrap

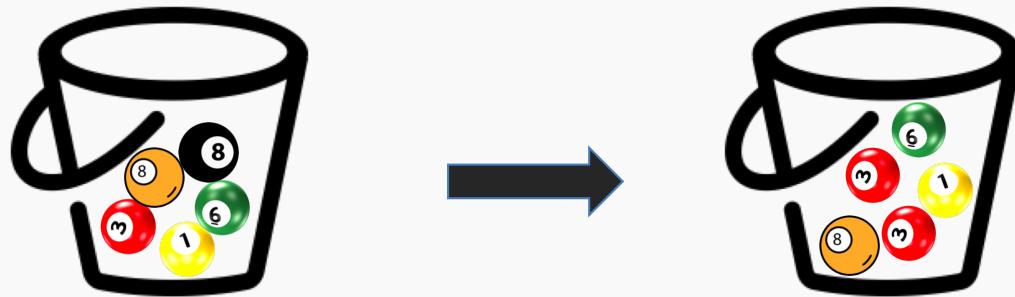
---

We repeat the same process and acquire another sample.



# Bootstrap

We repeat the same process and acquire another sample.



These new buckets represents the parallel universes

*iterations  $\approx 1000$*



# Bootstrapping for Estimating Sampling Error

## Definition

Bootstrapping is the practice of estimating properties of an estimator by measuring those properties by, for example, resampling from the observed data.

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  $\hat{\beta}_1$ .



## Confidence intervals for the predictors estimates: Standard Errors

We can empirically estimate the **standard errors**,  $SE(\hat{\beta}_0), SE(\hat{\beta}_1)$  of  $\hat{\beta}_0$  and  $\hat{\beta}_1$  through bootstrapping.

If for each bootstrapped sample the estimated betas are:  $\hat{\beta}_{0,i}, \hat{\beta}_{1,i}$ , then

$$SE(\hat{\beta}_0) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_{0,i})}$$
$$SE(\hat{\beta}_1) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_{1,i})}$$

Should be  
the same  
as what  
probability theory  
says it  
should



## Confidence intervals for the predictors estimates: Standard Errors

Alternatively:

If we know the variance  $\sigma_\epsilon^2$  of the noise  $\epsilon$ , we can compute  $SE(\hat{\beta}_0), SE(\hat{\beta}_1)$  analytically using the formulae below (no need to bootstrap):

$$SE(\hat{\beta}_0) = \sigma_\epsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}}$$
$$SE(\hat{\beta}_1) = \frac{\sigma_\epsilon}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

Let the bootstrap  
do this  
calculation for  
you

\*Note: these formulas do not rely on the normality assumption. But they do rely on independence, constant variance, and linearity.



## Standard Errors

$$SE(\hat{\beta}_0) = \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}}$$
$$SE(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$



**More data:**  $n \uparrow$  and  $\sum_i (x_i - \bar{x})^2 \uparrow \Rightarrow SE \downarrow$

**Larger coverage:**  $var(x)$  or  $\sum_i (x_i - \bar{x})^2 \uparrow \Rightarrow SE \downarrow$

**Better data:**  $\sigma^2 \downarrow \Rightarrow SE \downarrow$

In practice, we do not know the theoretical value of  $\sigma$  since we do not know the exact distribution of the noise  $\epsilon$ .



## Standard Errors

However, if we make the following assumptions,

- the errors  $\epsilon_i = y_i - \hat{y}_i$  and  $\epsilon_j = y_j - \hat{y}_j$  are uncorrelated, for  $i \neq j$ ,
- each  $\epsilon_i$  has a mean 0 and variance  $\sigma_\epsilon^2$ ,

then, we can empirically estimate  $\sigma^2$ , from the data and our regression line:

$$\hat{\sigma}^2 = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$$

In linear regression

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

Remember:

$$y_i = f(x_i) + \epsilon_i \Rightarrow \epsilon_i = y_i - f(x_i)$$



## Standard Errors

---

**More data:**  $n \uparrow$  and  $\sum_i(x_i - \bar{x})^2 \uparrow \Rightarrow SE \downarrow$

**Larger coverage:**  $var(x)$  or  $\sum_i(x_i - \bar{x})^2 \uparrow \Rightarrow SE \downarrow$

**Better data:**  $\sigma^2 \downarrow \Rightarrow SE \downarrow$

$$SE(\hat{\beta}_0) = \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}}$$
$$SE(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

---

**Better model:**  $(\hat{f} - y_i) \downarrow \Rightarrow \sigma \downarrow \Rightarrow SE \downarrow$

$$\hat{\sigma}^2 = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$$

**Question:** What happens to the  $\widehat{\beta}_0$ ,  $\widehat{\beta}_1$  under these scenarios?



## Standard Errors

The following results are for the coefficients for TV advertising:

Method	$SE(\hat{\beta}_1)$
Analytic Formula	0.0061
Bootstrap	0.0061

The coefficients for TV advertising but restricting the coverage of x are:

Method	$SE(\hat{\beta}_1)$
Analytic Formula	0.0068
Bootstrap	0.0068

The coefficients for TV advertising but with added **extra noise**:

Method	$SE(\hat{\beta}_1)$
Analytic Formula	0.0028
Bootstrap	0.0023

Does this make sense?

Due to randomness in the bootstrap

(maybe we didn't use enough resamples)



# Lecture Outline

---

Bootstrap Resampling

**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

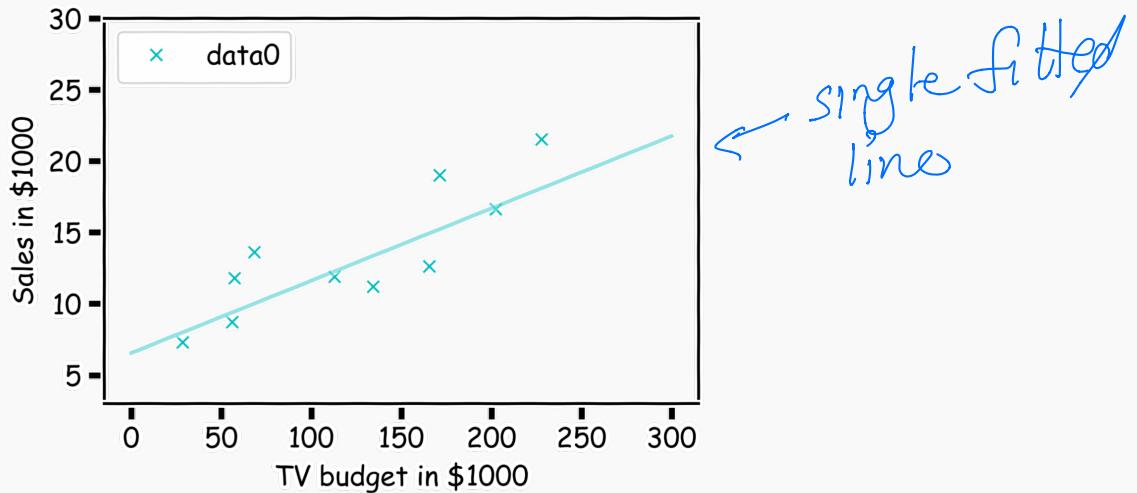
Variable Selection

Bias vs. Variance



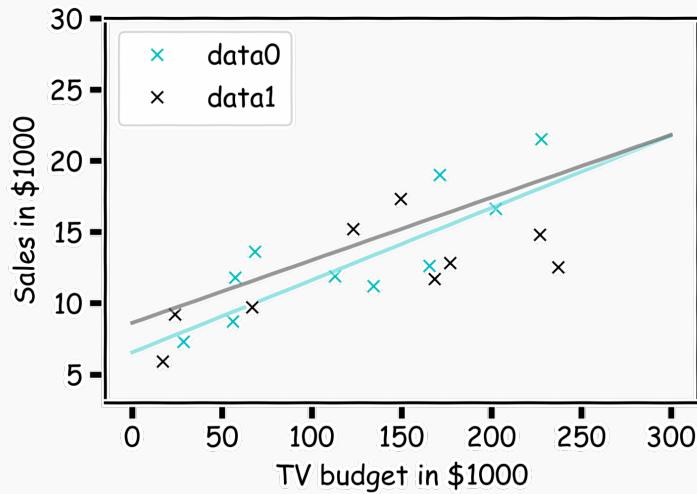
## How well do we know $\hat{f}$ ?

Our confidence in  $f$  is directly connected with the confidence in  $\beta$ s. So for each bootstrap sample, we have one  $\beta_0, \beta_1$  which we can use to predict  $y$  for all  $x$ 's.



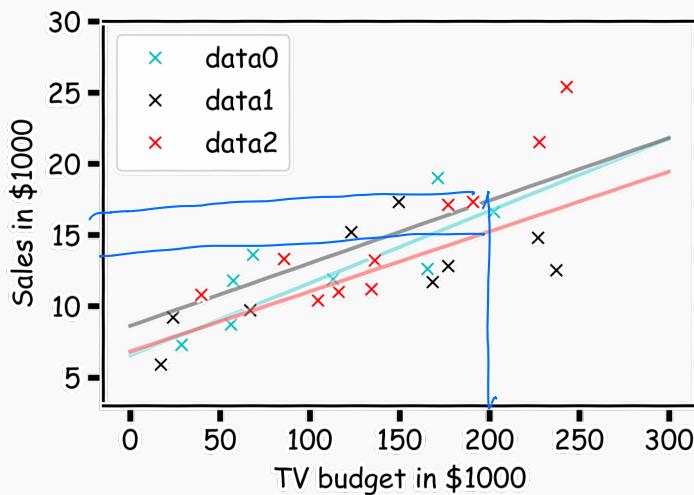
# How well do we know $\hat{f}$ ?

Here we show two different sets of models given the fitted coefficients.



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

There is one such regression line for every bootstrapped sample.



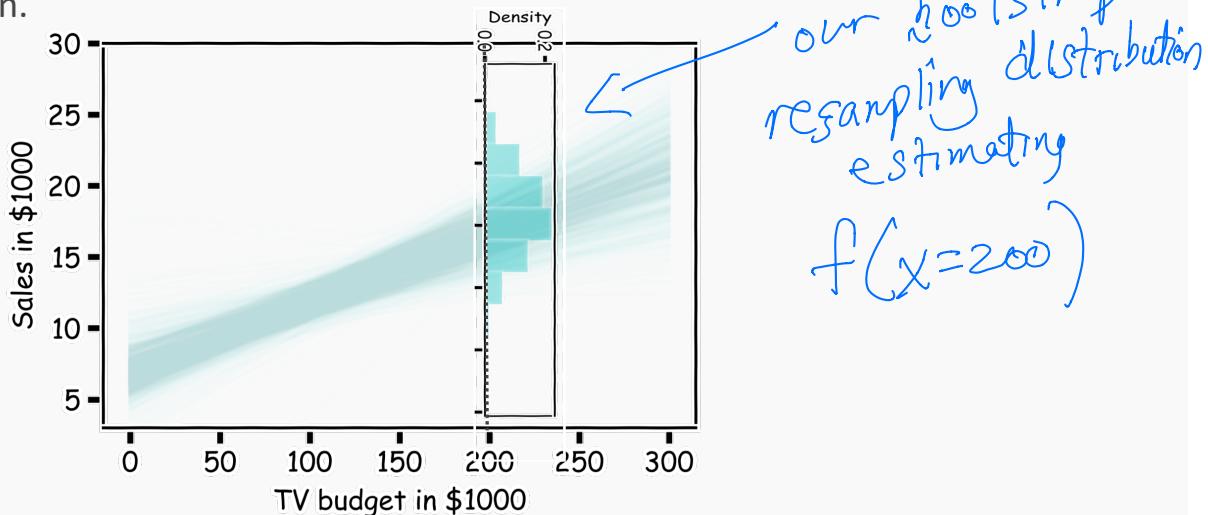
$x=200$   
 $\hat{f}(x=200) =$   
anywhere  
between 14  
17  
Based on parallel  
universe  
samples.



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

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

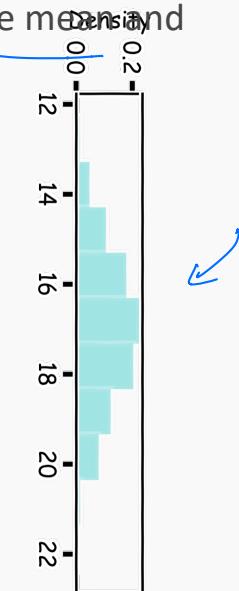
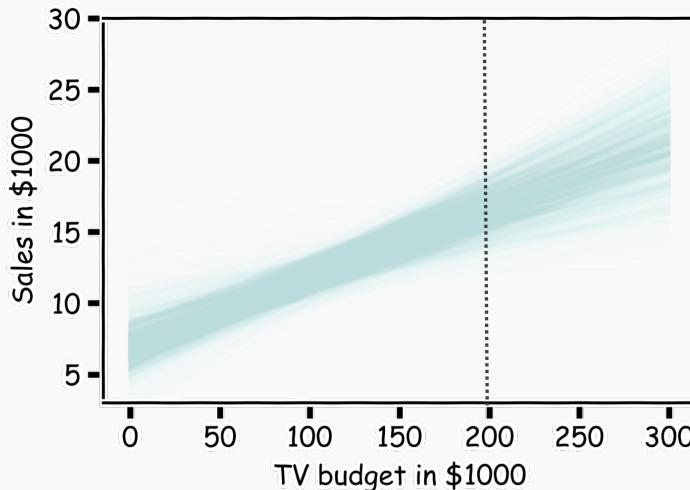
For a given  $x$ , we examine the distribution of  $\hat{f}$ , and determine the mean and standard deviation.



## 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.

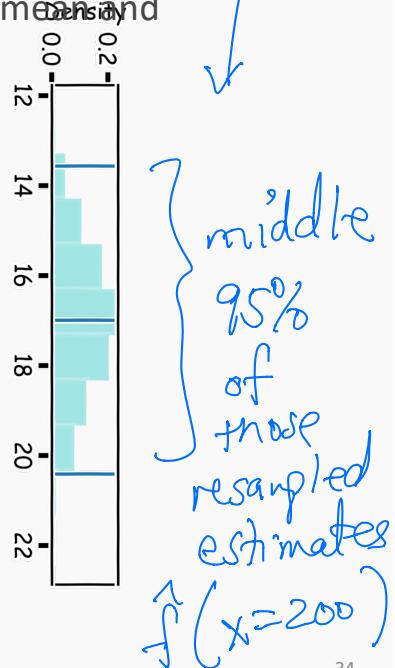
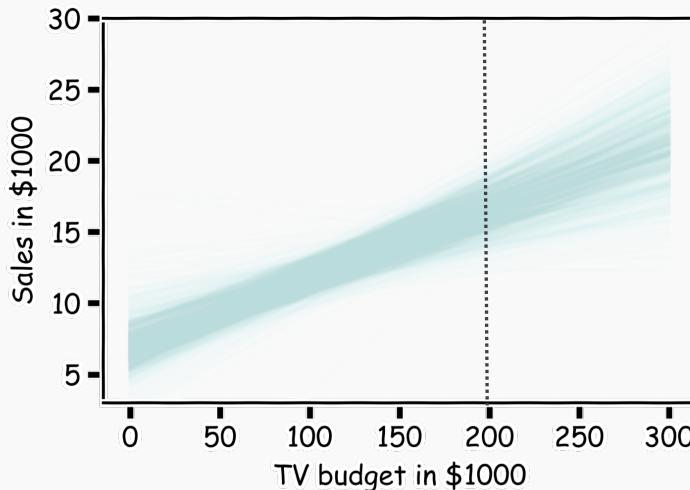


2.5% and  
97.5% percentiles

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

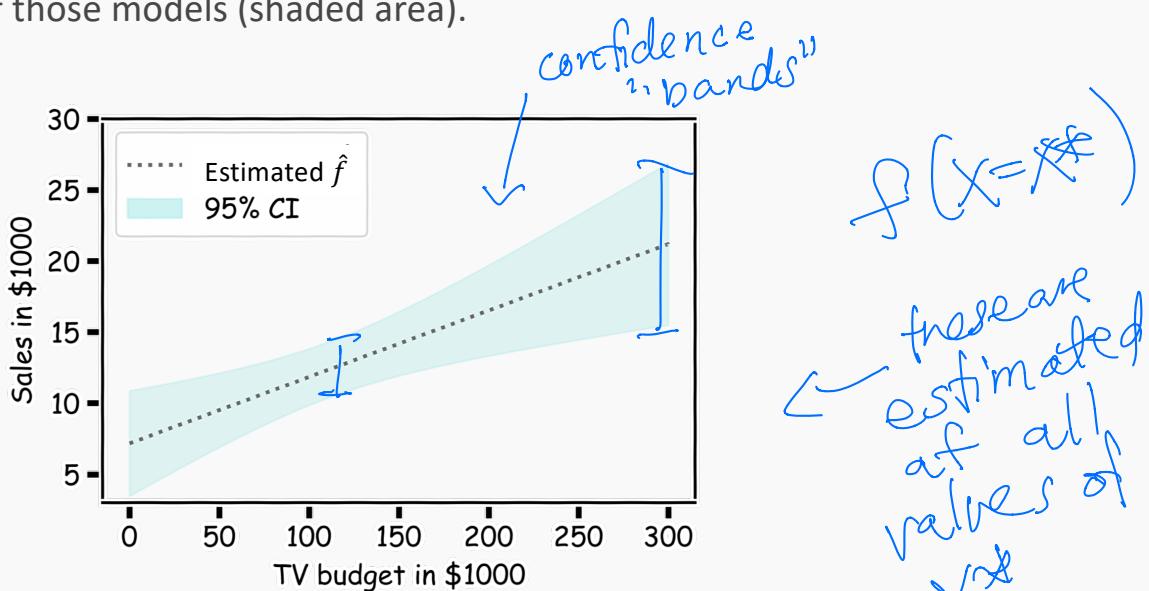
Below we show all regression lines for a thousand of such sub-samples.

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



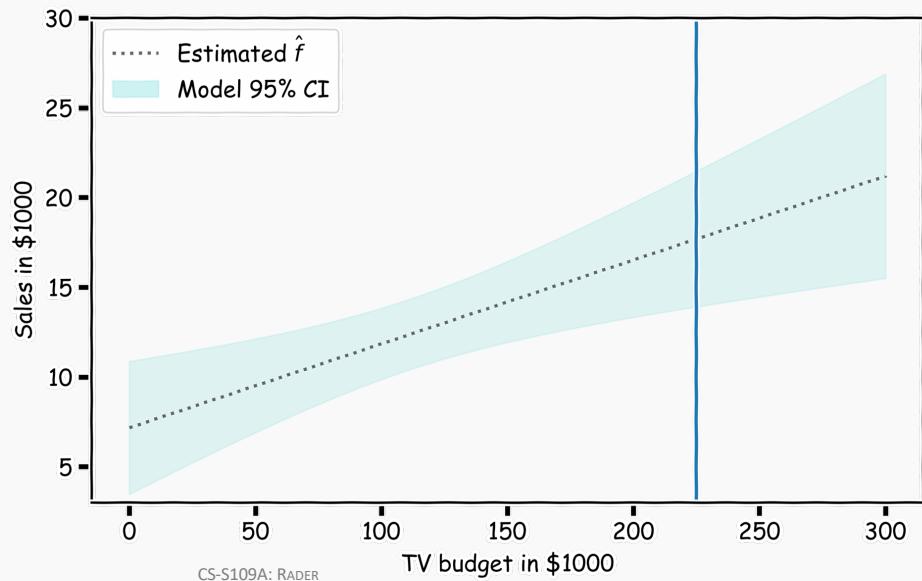
## How well do we know $\hat{f}$ ?

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).



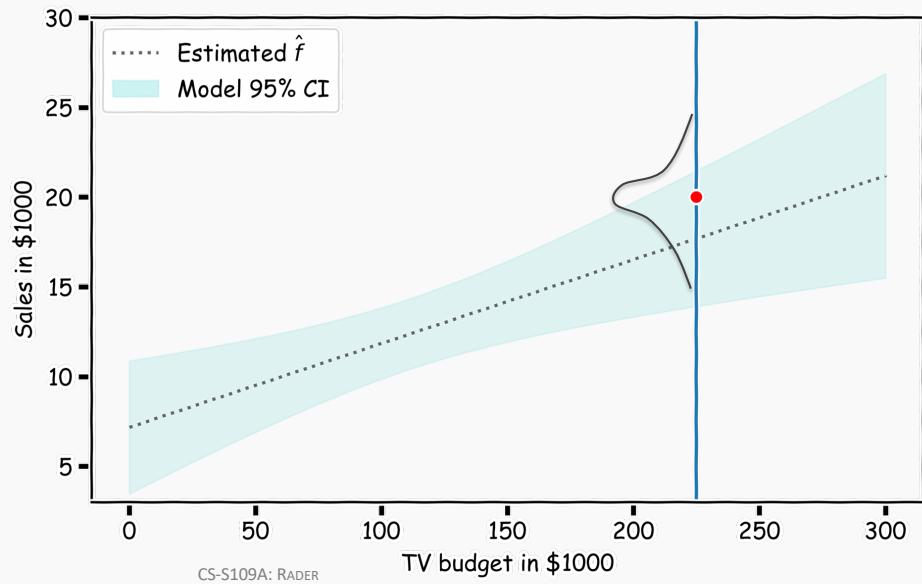
## Confidence in predicting $\hat{y}$

2nd quantity i) confidence interval of  $f(x=200)$   
ii) prediction interval of  $(\hat{y})$  at  $x=200$ .



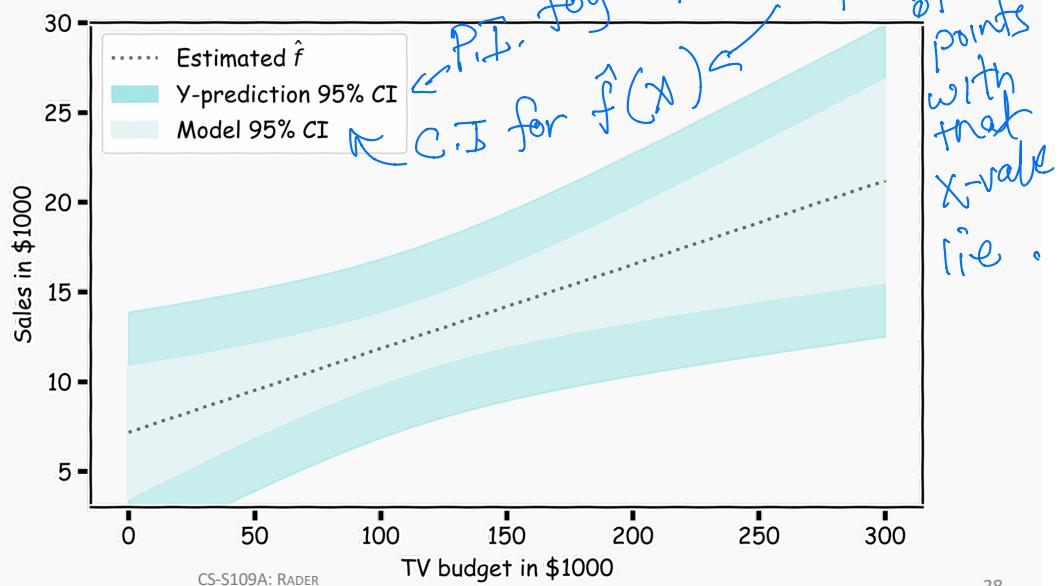
## Confidence in predicting $\hat{y}$

- 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

Bootstrap Resampling

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

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

} inference

## 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

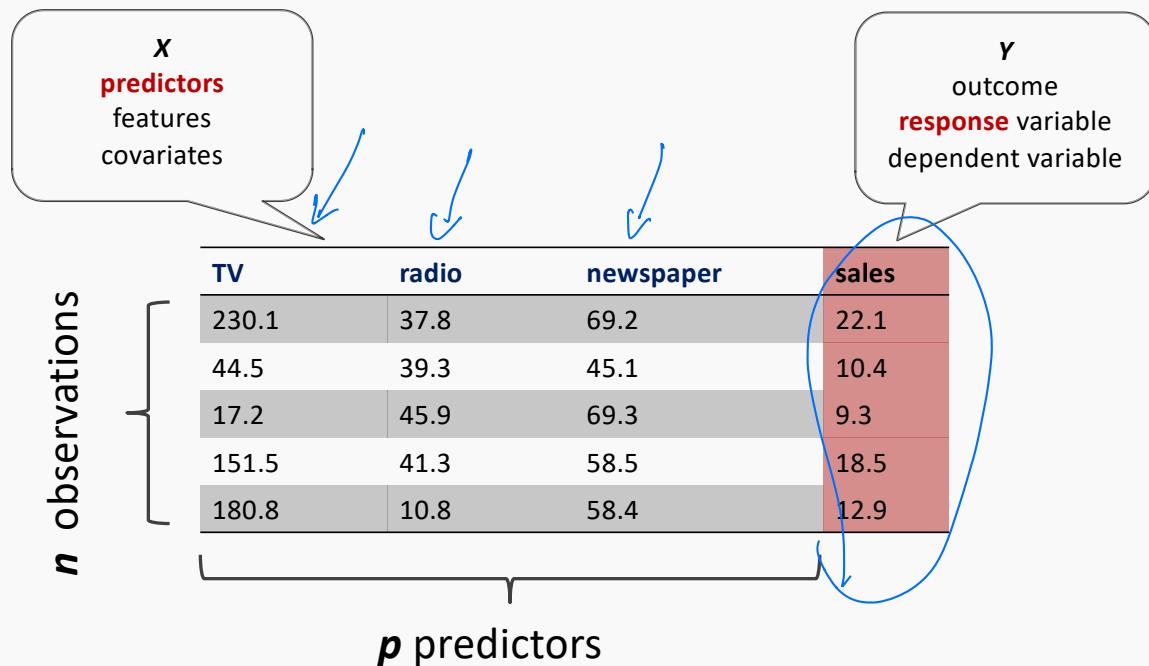
- Their weight, only
- Their weight and gender
- Their weight, gender, and income
- Their weight, gender, income, and favorite number

bringing in  
useless predictors  
that could "muddy"  
up the predictions  
from the model

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



## Multilinear Models

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

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

mathematically:  
plane or  
hyperplane

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

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

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

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



$(\hat{\beta}_1, \hat{\beta}_2, \dots)$  is estimated from data

## Multiple Linear Regression

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

Given a set of observations,

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

$$\vec{Y} = X\vec{\beta}$$

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

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



$\vec{Y}_{n \times 1}$  vector

CS-S109A: RADER  
to represent the intercept

$X$  matrix  
 $n \times (J+1)$

$\vec{\beta}_{(J+1) \times 1}$  vector

## Multilinear Model, example

For our data

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

X : "Design Matrix"

In linear algebra notation

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

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



# Multiple Linear Regression

The model takes a simple algebraic form:

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

noise around  
the hyperplane

Thus, the MSE can be expressed in vector notation as

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

Minimizing the MSE using vector calculus yields,

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \underset{\beta}{\operatorname{argmin}} \text{MSE}(\beta).$$

projection of  
 $\mathbf{Y}$  onto the  
column space of  
 $\mathbf{X}$ .

$$\begin{aligned} \mathbf{y}_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \\ &\quad + \beta_3 x_{i3} \dots \\ &\quad + \beta_j x_{ij} \end{aligned}$$



## Standard Errors for Multiple Linear Regression

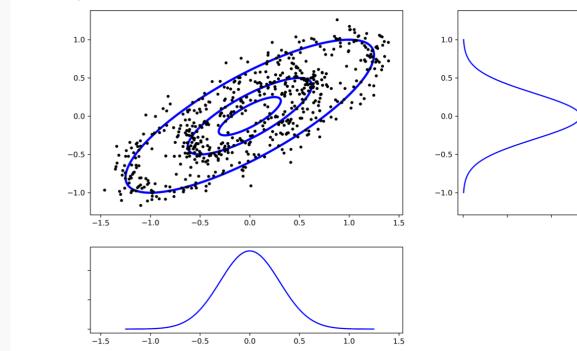
As with the simple linear regression, the standard errors can be calculated either using statistical modeling

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

Variance matrix of coefficient estimates

Alternatively, use the **bootstrap**.

? ↗



## Collinearity

linearly related

**Collinearity** 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.

\* when two columns in your  $X$  design matrix are correlated, then  $(X^T X)^{-1}$  is close to singular (hard to invert computationally)



what is the interpretation of  $\hat{\beta}_{\text{news}} = 0.013$ ?

## Collinearity

↳ the estimated change in Sales for

1-unit change in news advertising

One model

spending, when controlling for  
the other 2 predictors in  
the model.

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

	Coef.	Std.Err.	t	P> t	[0.025	0.975]
$\beta_0$	2.602	0.332	7.820	3.176e-13	1.945	3.258
$\beta_{\text{TV}}$	0.046	0.0015	29.887	6.314e-75	0.043	0.049
$\beta_{\text{RADIO}}$	0.175	0.0094	18.576	4.297e-45	0.156	0.194
$\beta_{\text{NEWS}}$	0.013	0.028	2.338	0.0203	0.008	0.035

→ changes a lot: is due to collinearity



Confounded with the effect of other 2 predictors

## Finding Significant Predictors: Hypothesis Testing

global test for usefulness of a model

For checking the significance of linear regression coefficients:

1. we set up our hypotheses  $H_0$ :

$$\rightarrow H_0 : \beta_0 = \beta_1 = \dots = \beta_J = 0$$

(Null)

$$H_1 : \beta_j \neq 0 \text{ for at least one } j$$

(Alternative)

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

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

$$F > 1 \Rightarrow \text{reject } H_0$$

CS-S109A: RADER

essentially testing  
if  $R^2$  is  
significantly greater  
than zero.



## Finding Significant Predictors: Hypothesis Testing

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

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

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

$\iff (p\text{-value} < 0.05?)$

$\uparrow \text{statsmodel will do the work for us.}$



# Lecture Outline

---

Bootstrap Resampling

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 (categorical)

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

Caucasian, Asian, Other



## 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:

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

"female"

$\beta_1$  = averaged difference  
comparing females to males

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

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

if  $x_i = 0$

$\beta_0$  = average for males



## Qualitative Predictors

---

**Question:** What is interpretation of  $\beta_0$  and  $\beta_1$ ?



## Qualitative Predictors

**Question:** What is interpretation of  $\beta_0$  and  $\beta_1$ ?

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

**Example:** Calculate  $\beta_0$  and  $\beta_1$  for the Credit data.

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

$$\begin{aligned}\hat{Y}_F &= 528 \\ &= \hat{\beta}_0 + \hat{\beta}_1\end{aligned}$$



$p = \# \text{ of categories}$

$(p-1)$   
dummies

## 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:

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

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

2 binary / indicators  
fully define  
fine 3  
unique groups



## More than two levels: One hot encoding

pandas  
sklearn  
do this  
for us  
"get\_dummies"

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  $\beta_0, \beta_1, \beta_2$ ?

estimated difference comparing Asians to African-Americans  
estimated average for the reference group (African-Americans)



# Lecture Outline

---

Bootstrap Resampling

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**

"synergistic" or "dampening" effects

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  $\beta_1$ , regardless of the amount spent on radio.

**Synergy effect or interaction effect** states that when an increase on the radio budget affects the effectiveness of the TV spending on sales.



## Beyond linearity

We change

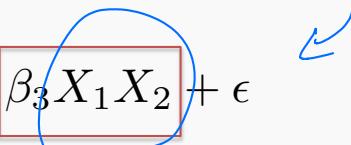
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

To

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

main  
effects

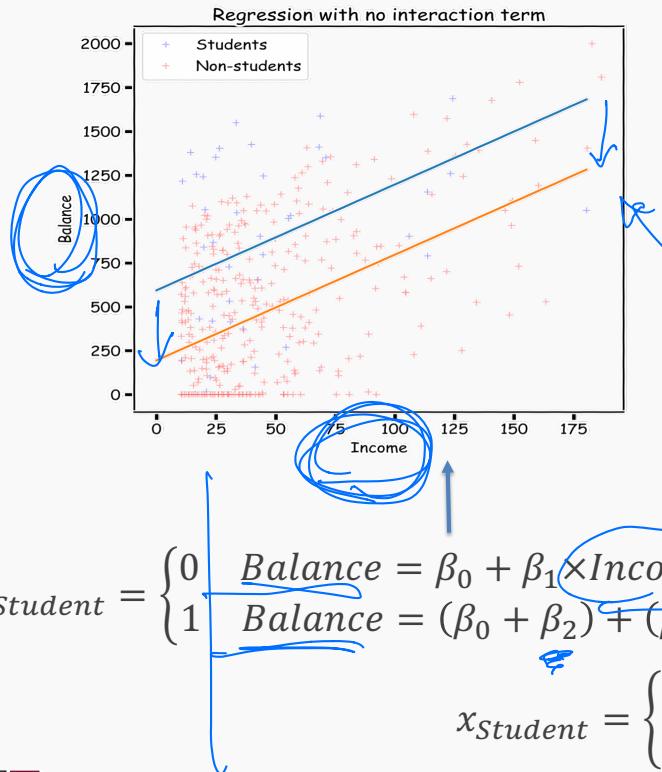
multiple regression  
with  $X_1$  &  $X_2$



mathematical  
multiplication  
of the 2 predictors  
together  
(as a 3rd variable  
in the model)



## What does it mean?



$x_1 = \text{income}$   
 $x_2 = \text{student}$  ← Binary indicators

allows these lines to be non-parallel

$\beta_3$ : measuring the difference in "effects" of Income on credit card balance comparing student to non-student

Effect of Income is potentially different for students



## Predictors predictors predictors

We have a lot predictors!

Is it a problem?

**Yes:** Computational Cost

**Yes:** Overfitting

Wait there is more ...

time to fit the model

affects both interpretations (inferences)  
and predictive ability  
(out-of-sample)



## 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

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

CS-S109A: RADER



single predictors

non-linear relationship

investigate the residuals here see if there is any non-linear relationship

63

$$r_i = e_i^* = y_i - \hat{y}_i$$

↑  
Residuals

## Residual Analysis

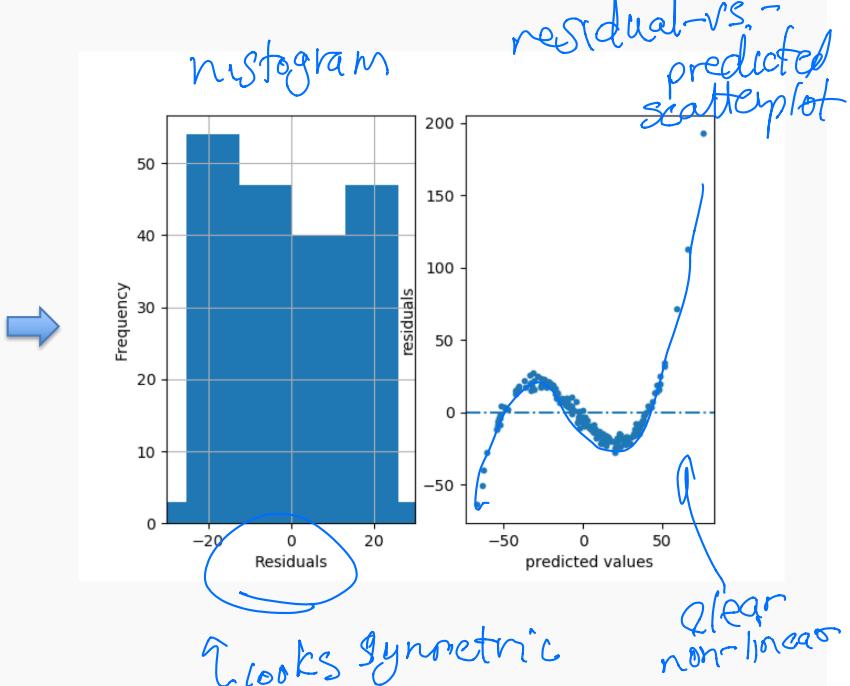
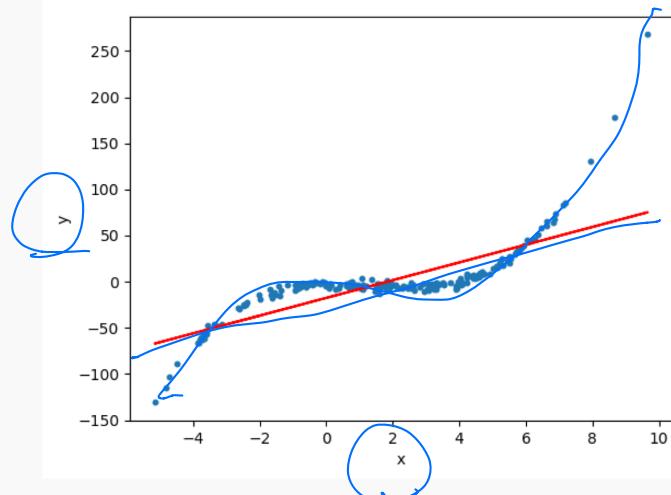
When we estimated the variance of  $\epsilon$ , we assumed that the residuals  $r_i = y_i - \hat{y}_i$  were uncorrelated and normally distributed with mean 0 and fixed variance.

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  $\hat{y}_i$ .
- (patterns in the residual scatterplots)*



# Residual Analysis



# Lecture Outline

---

Bootstrap Resampling

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

one way  
to model  
non-linearities



## Polynomial Regression

order -  $M$   
~~poly~~( $M$ )

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

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

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

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1^1 & \dots & x_1^M \\ 1 & x_2^1 & \dots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^1 & \dots & x_n^M \end{pmatrix}, \quad \text{columns represent the } M \text{ polynomial terms}$$

$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_M \end{pmatrix}.$



↑ multiple regression matrix version

## Polynomial Regression

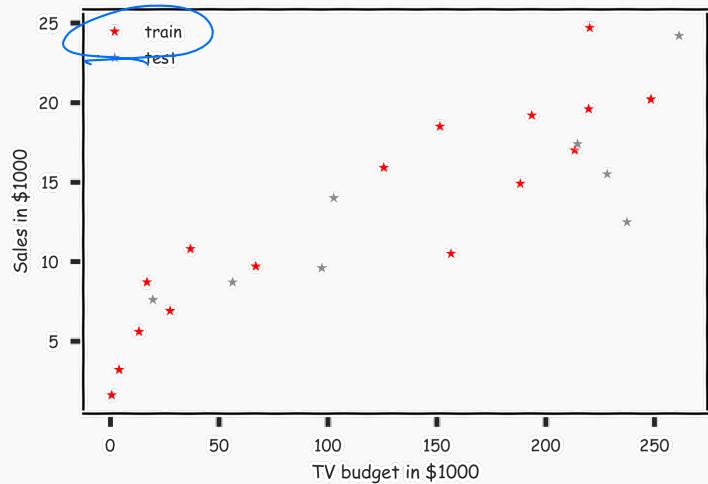
Again, minimizing the MSE using vector calculus yields,

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \text{MSE}(\beta) = \left( \mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{Y}.$$

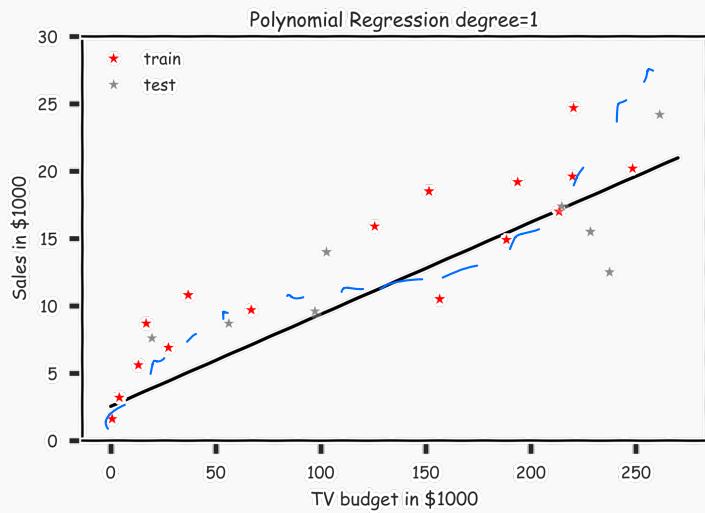
↑ by definition: the columns of  $\mathbf{X}$  are likely very collinear



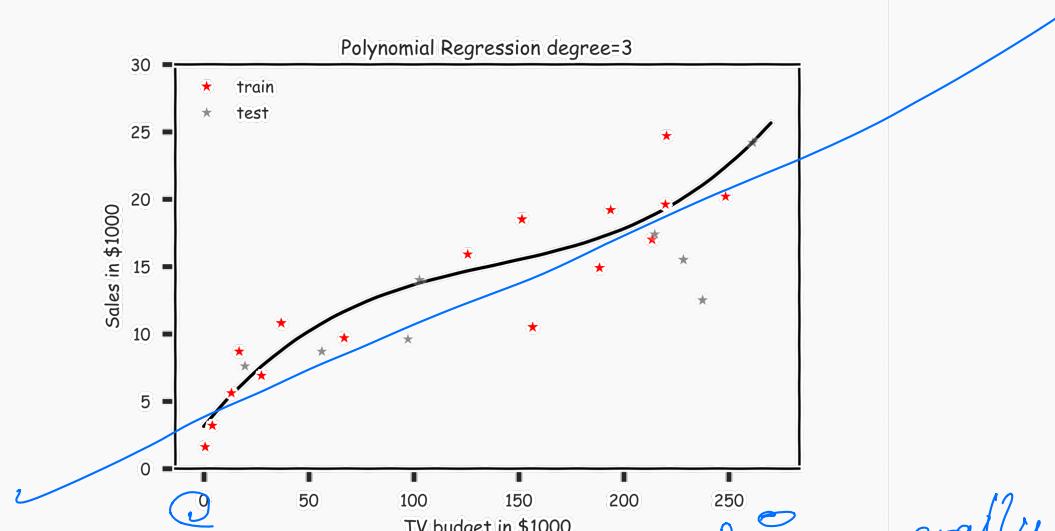
## Polynomial Regression (cont)



## Polynomial Regression (cont)

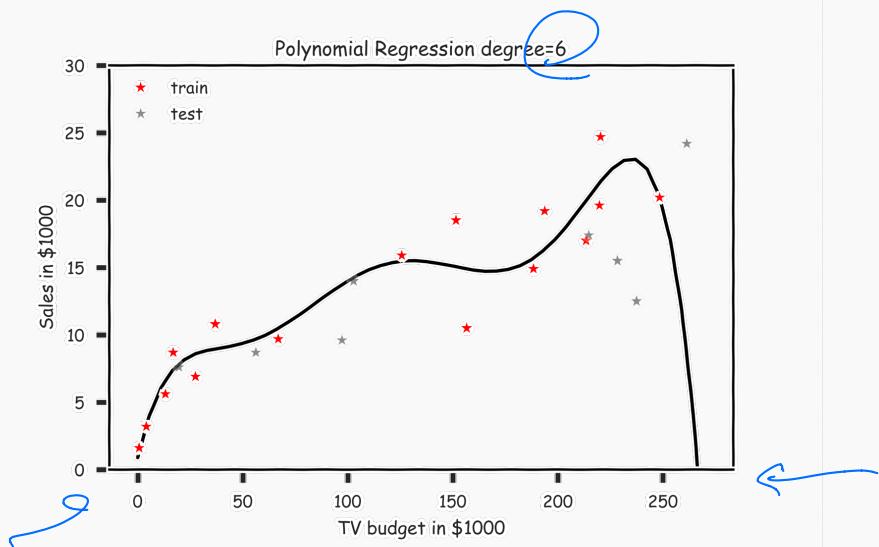


## Polynomial Regression (cont)

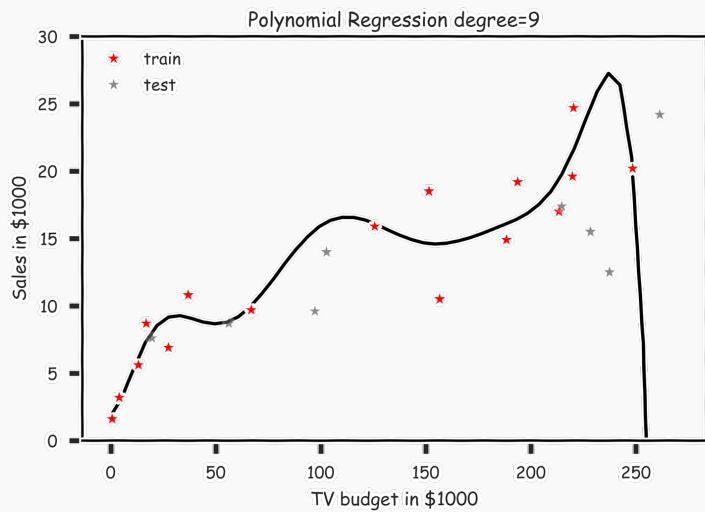


odd order polynomials will give generally patterns  
of increase or decrease

## Polynomial Regression (cont)



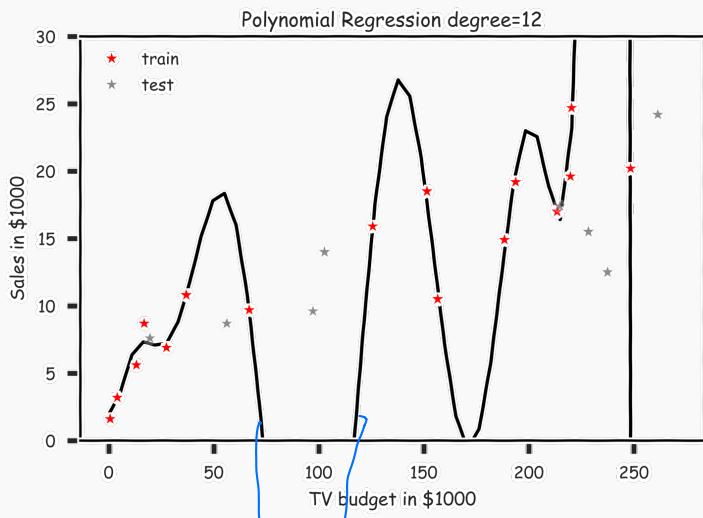
## Polynomial Regression (cont)



## Polynomial Regression (cont)

$n=15$  in train

$$|\beta| = 13$$



Illustrating overfitting to the training data

CS-S109A: RADER

$$\hat{f}(x=100) = -10$$



## Polynomial Regression

linear algebra  
set-up

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

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

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

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



→ no such thing as polynomial predictors  
for dummies / indicators

## 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

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



polynomial of  $X_1$   
along with  $X_2, X_3, \dots, X_J$   
polynomials, or interactions, etc...

# Lecture Outline

---

Bootstrap Resampling

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

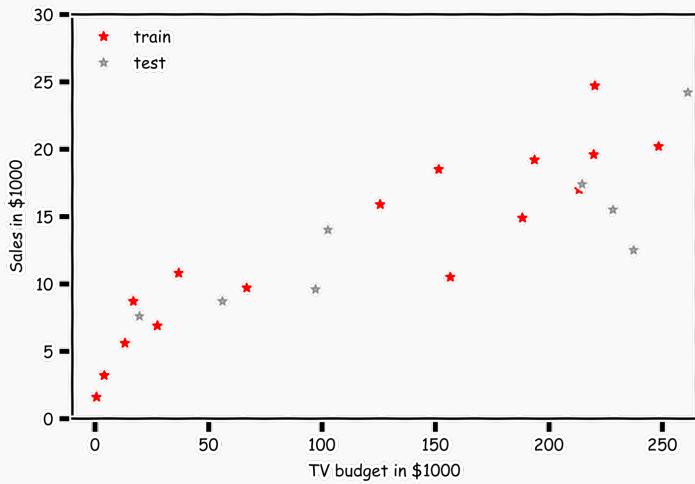
## **Overfitting**

Variable Selection

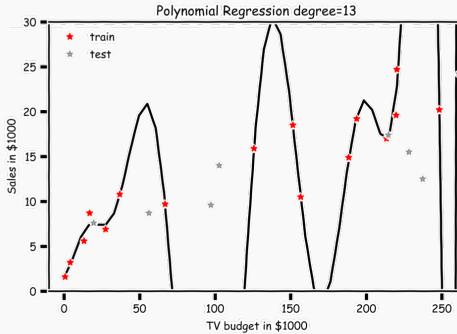
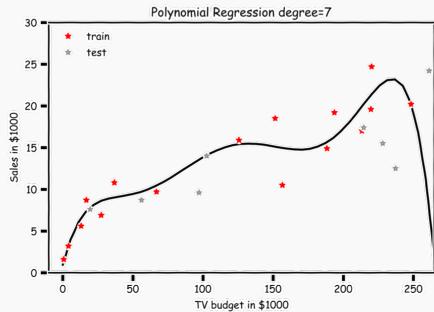
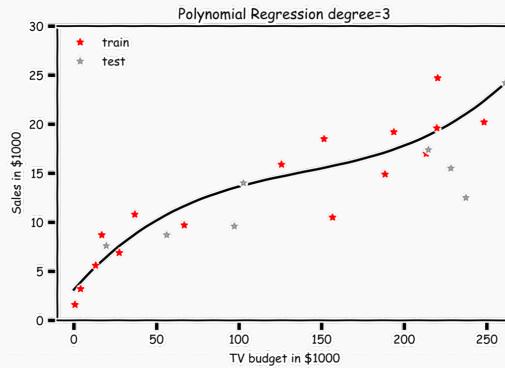
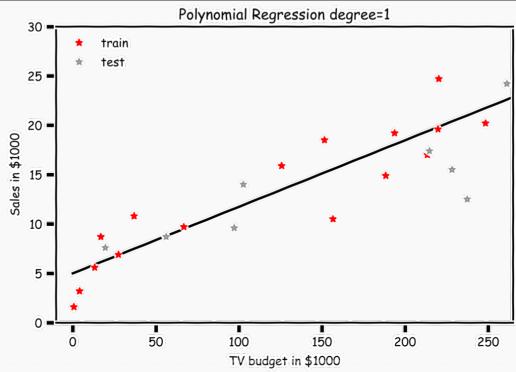
Bias vs. Variance



# Overfitting

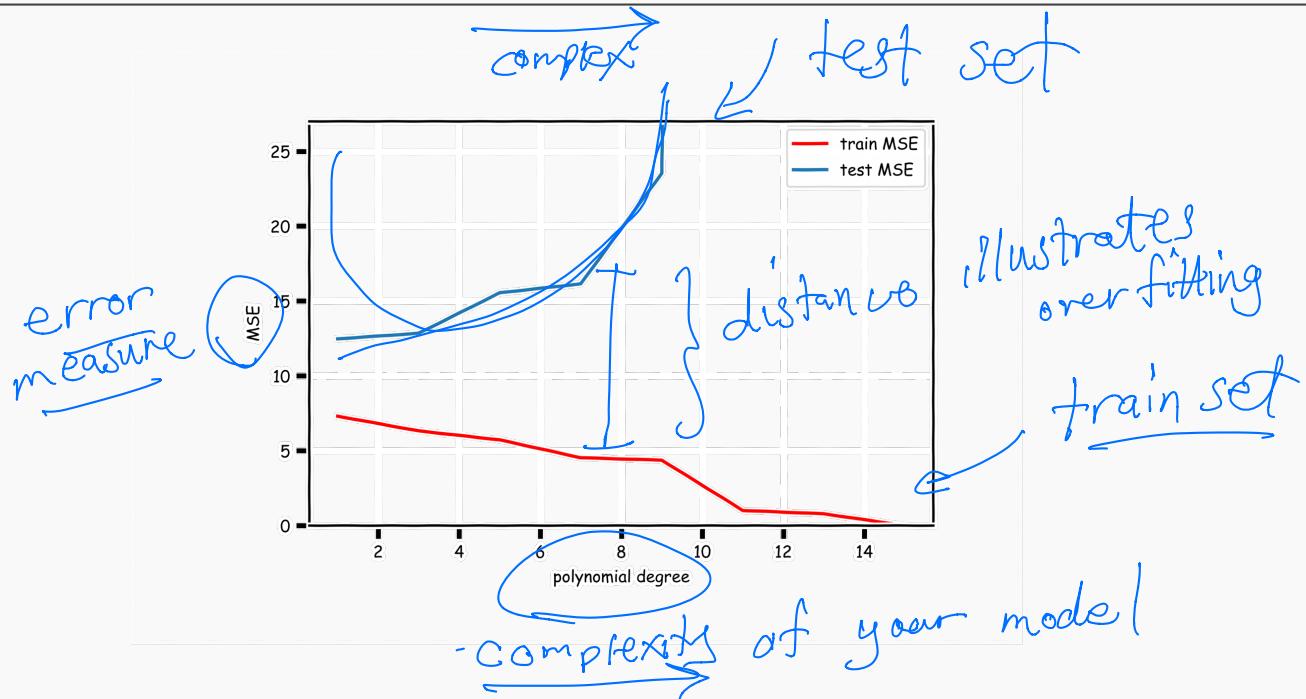
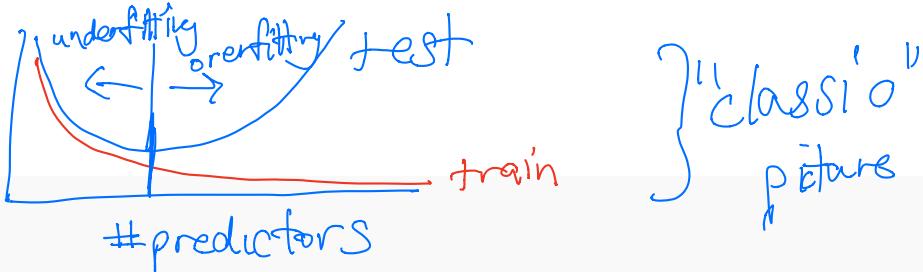


# Overfitting with polynomials



## MSE

### 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:

*begins to fit to the particular noise in the training set.*

Think:

$$Y = f(X) + \varepsilon$$



## Train-Validation-Test

### Question:

What is a more honest representation of the performance of the model?

$$\begin{aligned} R^2_{\text{test}}(\text{degree}=1) &= 0.52 \\ R^2_{\text{train}}(\text{degree}=1) &= 0.83 \end{aligned}$$

} indication of overfitting  
(they are very different)

Always use out-of-sampling measures of accuracy or performance



# Lecture Outline

---

Bootstrap Resampling

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?

2<sup>J</sup>



# Model Selection

## Example: 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 and 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,
- validation - selecting an optimal model by evaluating each model on validation/test set.  
*or out-of-sample*

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  $P_0$ , construct the null model  $M_0$ .

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

    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  $\{M_0, M_1, \dots, M_J\}$  that optimizes a fixed metric (this can be validation MSE,  $R^2$ ; or AIC/BIC on training set)

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

Backward  
Both ways



Adv) quicker, generally  
Dis) prone to multicollinearity

## 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

---

Bootstrap Resampling

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

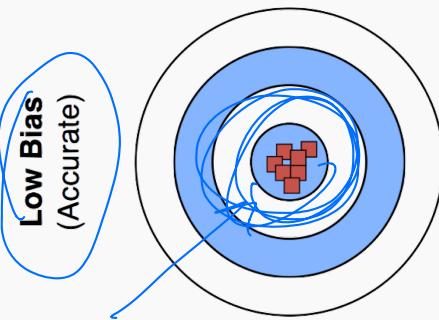
Variable Selection

**Bias vs. Variance**

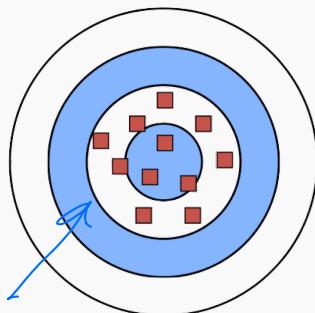
*Trade-off*



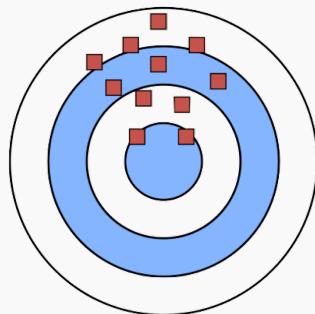
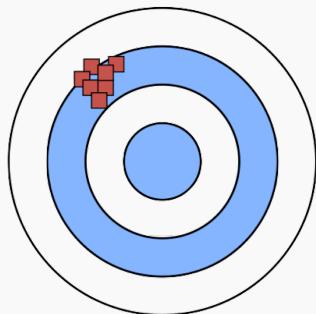
**Low Variance**  
(Precise)



**High Variance**  
(Not Precise)



**High Bias**  
(Not Accurate)



CS-S109A: RADER

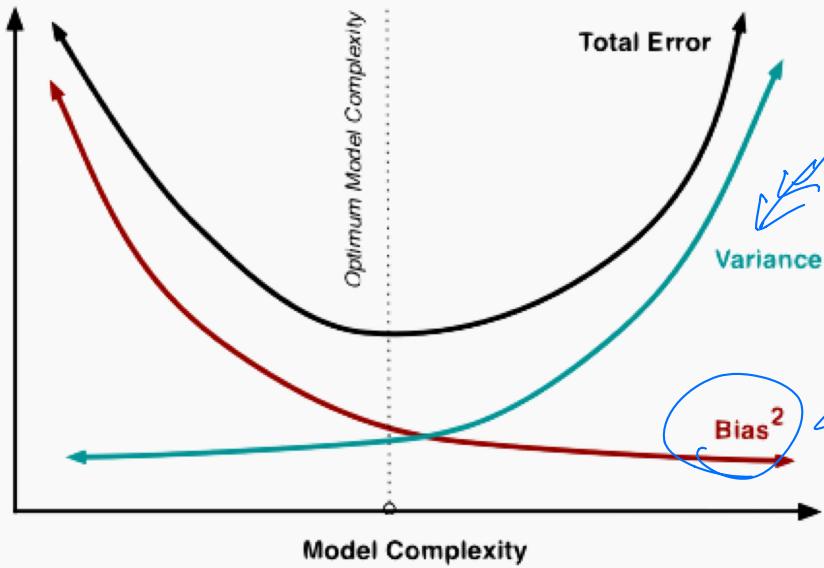


↑ on average, very  
spread out

MSE in test

Out-of-sample  
Prediction Error

Error



Bias<sup>2</sup>

estimates of  $f$  will vary wildly if model is complex

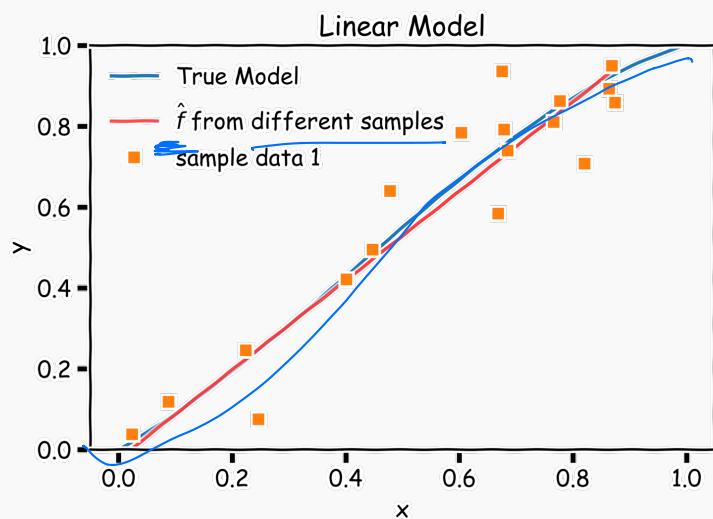
more complex models have the potential to capture the true "f"



$$MSE = \underbrace{\text{variance}}_{\text{CS-S109A: RADER}} + \underbrace{\text{Bias}^2}_{\text{f(x)}}}$$

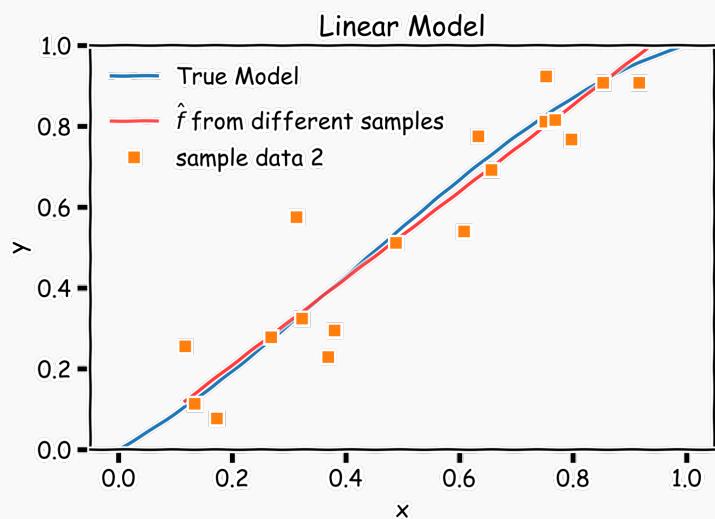
} variance and bias of using  $\hat{f}(x)$  to estimate  $f(x)$

## Bias vs Variance

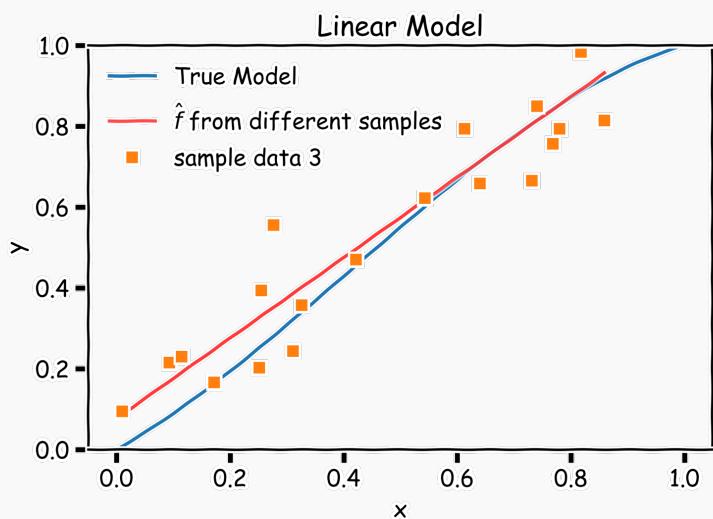


Truth: a little  
non-linear

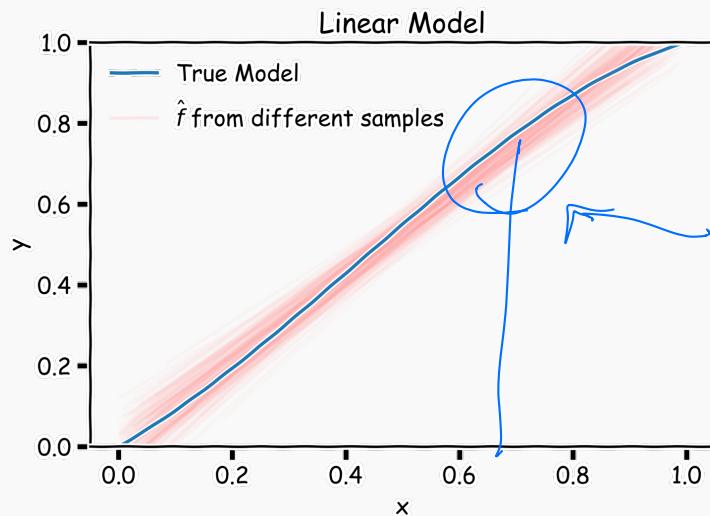
## Bias vs Variance



## Bias vs Variance



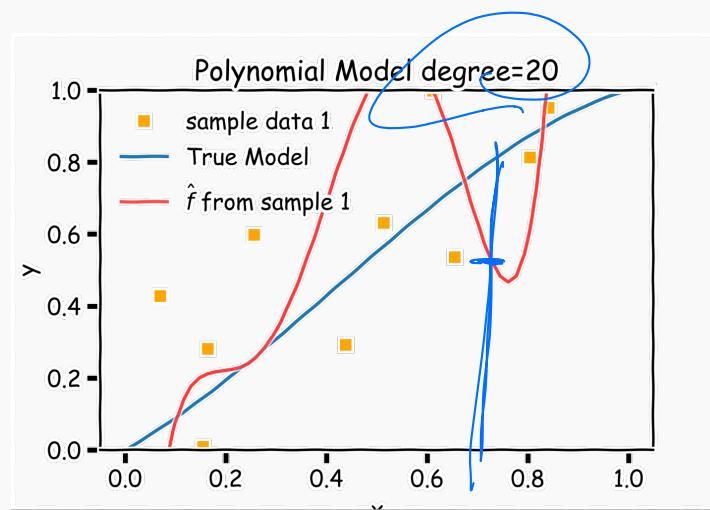
## Linear models: 20 data points per line 2000 simulations



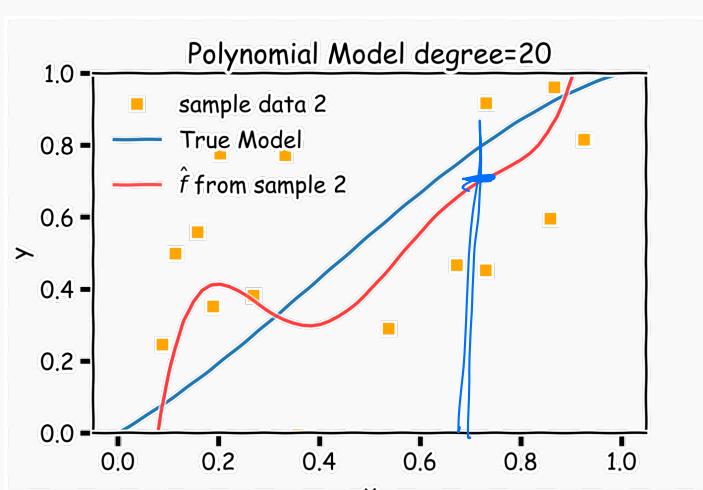
off a little here  
(underestimating  
truth, on  
average)



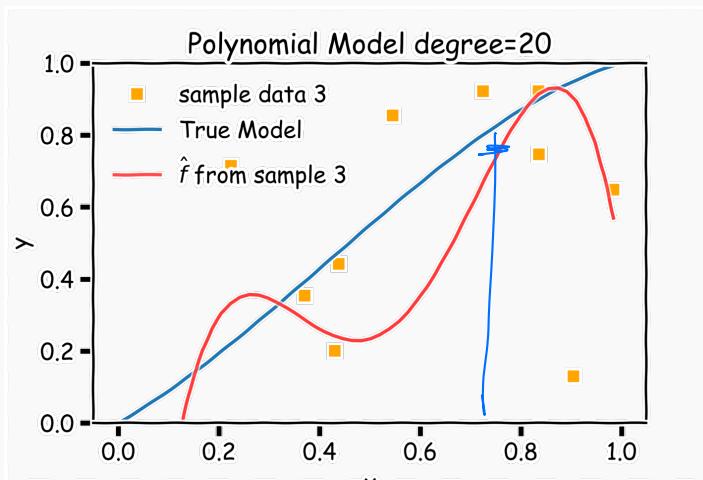
## Bias vs Variance



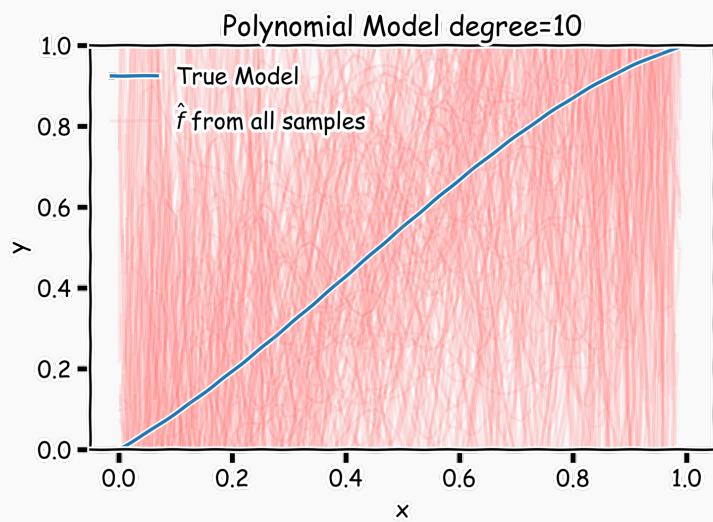
## Bias vs Variance



## Bias vs Variance



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



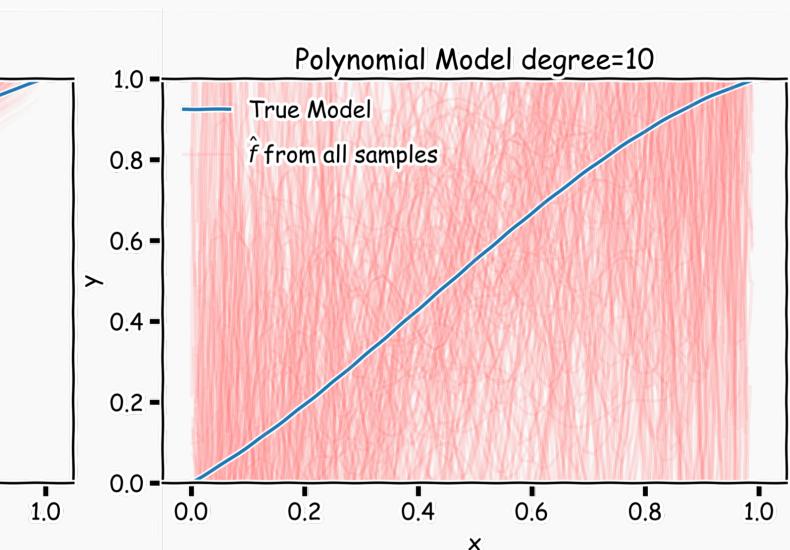
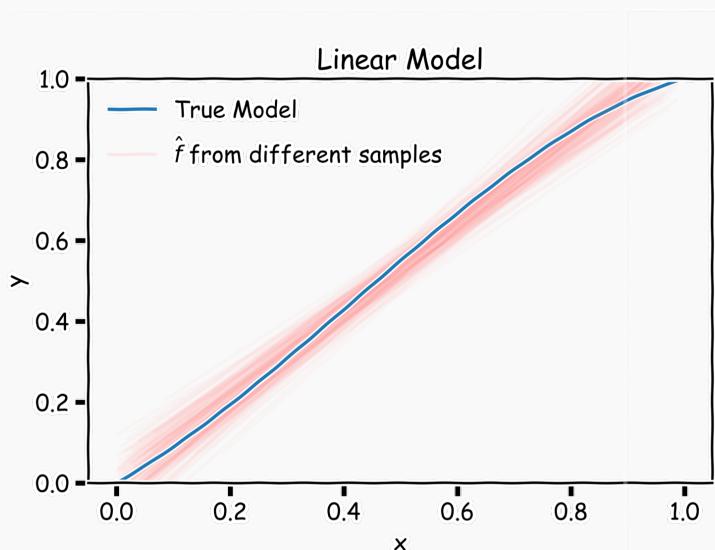
← they would likely  
on average be  
at the right  
place (blue  
line)  
the variability  
(S) is too  
overwhelming



## Bias vs Variance

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

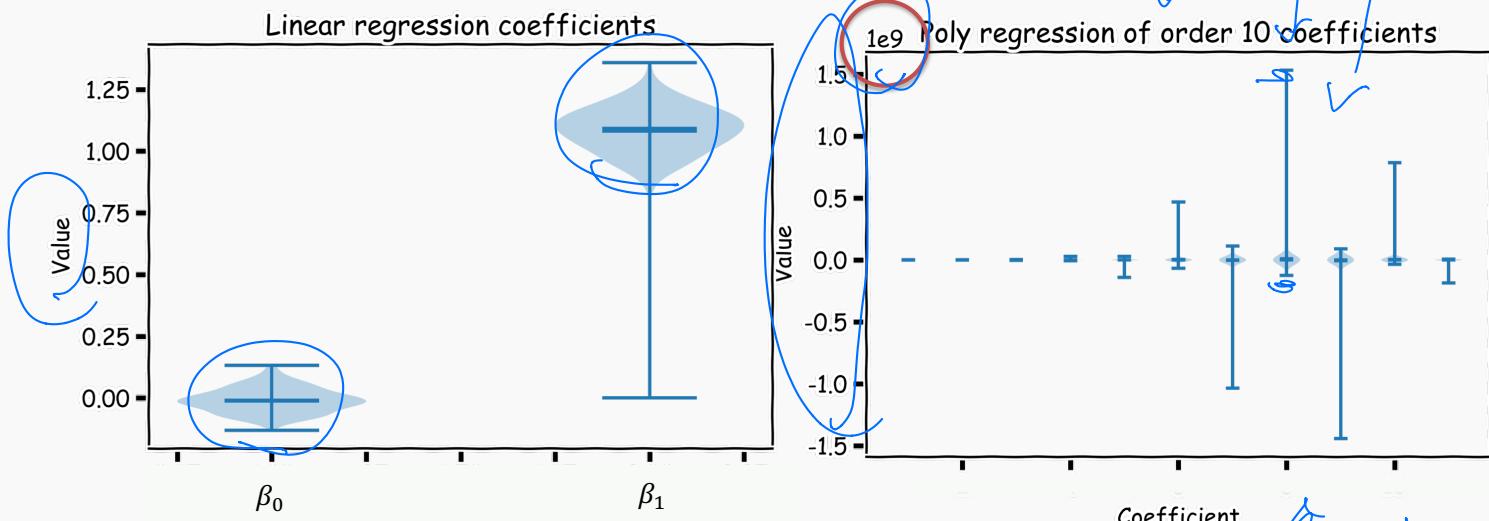
Right: Best-fit models using degree 10 polynomial



## Bias vs Variance

Left: Linear regression coefficients

Right: Poly regression of order 10 coefficients



↑ high multicollinearity  
⇒ overfitting