

# Ch 3.3: Even More Linear Regression

## Lecture 7 - CMSE 381

Prof. Mengsen Zhang

Michigan State University

::

Dept of Computational Mathematics, Science & Engineering

Wed, Jan 29, 2025

## **Last time:**

- 3.2 Multiple Linear Regression

## **Announcements:**

- HW #2 Due Sunday!
- Office hours

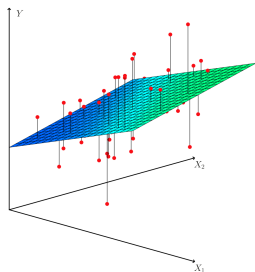
## Covered in this lecture

- RSE,  $R^2$
- Confidence intervals and prediction intervals
- Qualitative predictors

## Section 1

Continued: Questions to ask of your model

# Linear Regression with Multiple Variables



- Predict  $Y$  on a multiple variables  $X$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

- Find good guesses for  $\hat{\beta}_0, \hat{\beta}_1, \cdots$ .
- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \cdots + \hat{\beta}_p x_{ip}$

- $e_i = y_i - \hat{y}_i$  is the  $i$ th residual
- $RSS = \sum_i e_i^2$
- $RSS$  is minimized at *least squares coefficient estimates*

## Review: Questions to ask of your model

- ① Is at least one of the predictors  $X_1, \dots, X_p$  useful in predicting the response?
- ② Do all the predictors help to explain  $Y$ , or is only a subset of the predictors useful?

Q3

How well does the model fit the data?

# Assessing the accuracy of the module

Almost the same as before

**Residual standard error (RSE):**

$$RSE = \sqrt{\frac{1}{n - p - 1} RSS}$$

**R squared:**

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$TSS = \sum_i (y_i - \bar{y})^2$$



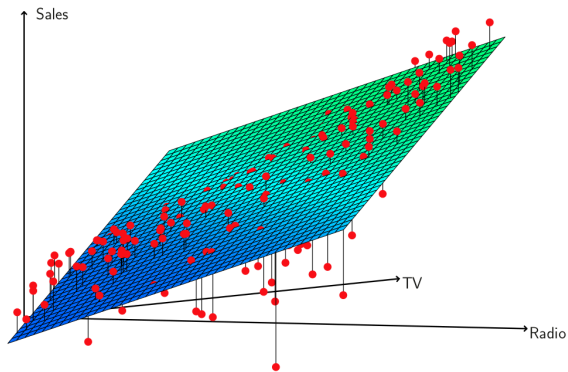
## $R^2$ on Advertising data

- Just TV:  $R^2 = 0.61$
- Just TV and radio:  $R^2 = 0.89719$
- All three variables:  $R^2 = 0.8972$

# RSE on Advertising Data

- Just TV:  $RSE = 3.26$
- Just TV and radio:  $RSE = 1.681$
- All three variables:  $RSE = 1.686$

# If all else fails, look at the data



Q4

Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

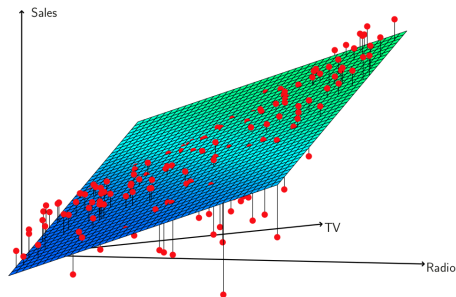
## Q4: Making predictions

Given estimates  $\hat{\beta}_0, \dots, \hat{\beta}_p$  for  $\beta_0, \dots, \beta_p$   
Least squares plane:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p$$

estimate for the true population regression plane

$$f(X) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$



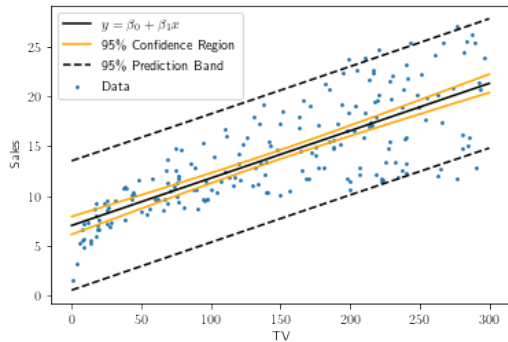
# Confidence vs Prediction Model

## Confidence Interval

The range likely to contain the population parameter (mean, standard deviation) of interest.

## Prediction Interval

The range that likely contains the value of the dependent variable for a single new observation given specific values of the independent variables.



## Specific to the Advertising Data

**Confidence interval:** quantify the uncertainty surrounding the average sales over a large number of cities.

**Advertising example:**

If \$100K is spent on TV, and \$20K on radio, **in each of  $n$  cities**

95% CI for average sales:  
[10,985, 11,528].

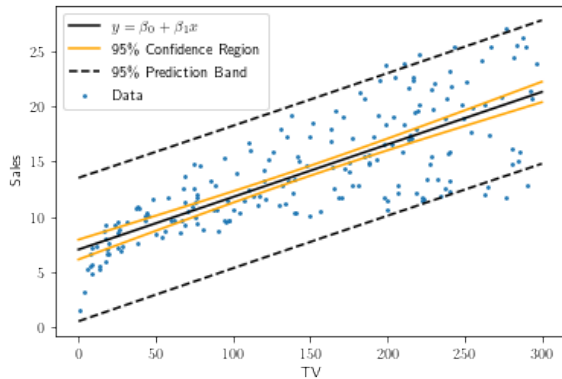
**Prediction Interval:** quantify the uncertainty in sales for a particular city.

**Advertising example:**

Given that \$100,000 is spent on TV advertising and \$20,000 is spent on radio advertising in **Gotham City**

95% prediction interval for Gotham:  
[7,930, 14,580].

# Comparing the two





Go take a look at the code under Q4

## Review: Questions to ask of your model

- 1 Is at least one of the predictors  $X_1, \dots, X_p$  useful in predicting the response?
- 2 Do all the predictors help to explain  $Y$ , or is only a subset of the predictors useful?
- 3 How well does the model fit the data?
- 4 Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

## Section 2

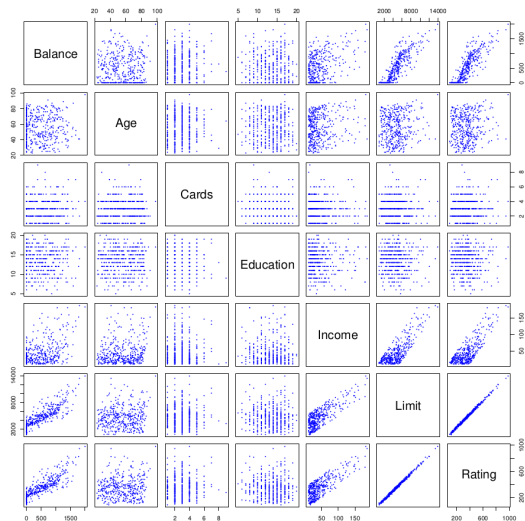
### Qualitative Predictors

## Reminder: Qualitative vs Quantitative predictors

**Quantitative:**

**Qualitative/Categorical:**

# New data set! Credit card balance



- own: house ownership
- student: student status
- status: marital status
- region: East, West, or South

# What if....

... your variables aren't quantitative?

- Home ownership
- Student status
- Major
- Gender
- Ethnicity
- Country of origin

## Example

Investigate differences in credit card balance between people who own a house and those who don't, ignoring the other variables.

# One-hot encoding

Create a new variable

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases}$$

Model:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i \\ &= \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is student} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person isn't} \end{cases} \end{aligned}$$

# Interpretation

	coef	std err	t	P> t	[0.025	0.975]
Intercept	480.3694	23.434	20.499	0.000	434.300	526.439
Student[T.Yes]	396.4556	74.104	5.350	0.000	250.771	542.140

Model:

$$y = 480.36 + 396.46 \cdot x_{student}$$



# Who cares about 0/1?

## Old version: 0/1

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases}$$

Model:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i \\ &= \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is student} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person isn't} \end{cases} \end{aligned}$$

## Alternative version: $\pm 1$

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is a student} \\ -1 & \text{if } i\text{th person is not a student} \end{cases}$$

Model:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i \\ &= \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is student} \\ \beta_0 - \beta_1 + \varepsilon_i & \text{if } i\text{th person isn't} \end{cases} \end{aligned}$$

# Qualitative Predictor with More than Two Levels

Region:

	$x_{i1}$	$x_{i2}$
South		
West		
East		

Create spare dummy variables:

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person from South} \\ 0 & \text{if } i\text{th person not from South} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person from West} \\ 0 & \text{if } i\text{th person not from West} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

$$= \begin{cases} \beta_0 + \beta_1 x_{i1} + \varepsilon_i & \text{if } i\text{th person from South} \\ \beta_0 + \beta_2 x_{i2} + \varepsilon_i & \text{if } i\text{th person from West} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person from East} \end{cases}$$

## More on multiple levels

	Coefficient	Std. error	<i>t</i> -statistic	<i>p</i> -value
Intercept	531.00	46.32	11.464	< 0.0001
region[South]	-18.69	65.02	-0.287	0.7740
region[West]	-12.50	56.68	-0.221	0.8260

Do code section on "Playing with multi-level variables"

# Next time

CMSE381\_S2025\_Schedule : Sheet1

Lec #	Date	Topic	Reading	HW
1	M 1/13	Intro / Python Review	1	
2	W 1/15	What is statistical learning	2.1	
3	F 1/17	Assessing Model Accuracy	2.2.1, 2.2.2	
	M 1/20	MLK - No Class		
4	W 1/22	Linear Regression	3.1	
5	F 1/24	More Linear Regression	3.1	HW #1 Due Sun 1/26
6	M 1/27	Multi-linear Regression	3.2	
7	W 1/29	Probably More Linear Regression	3.3	
8	F 1/31	Last of the Linear Regression		HW #2 Due Sun 2/1
9	M 2/3	Intro to classification, Bayes classifier, KNN classifier	2.2.3	
10	W 2/5	Logistic Regression	4.1, 4.2, 4.3.1-3	
11	F 2/7	Multiple Logistic Regression / Multinomial Logistic Regression	4.3.4-5	HW #3 Due Sun 2/9
	M 2/10	<b>Project Day &amp; Review</b>		
	W 2/12	<b>Midterm #1</b>		
12	F 2/14	Leave one out CV	5.1.1, 5.1.2	
13	M 2/17	k-fold CV	5.1.3	
14	W 2/19	More k-fold CV	5.1.4-5	
15	F 2/21	k-fold CV for classification	5.1.5	HW #4 Due Sun 2/23
16	M 2/24	Subset selection	6.1	