Ch 3.3: Even More Linear Regression

Lecture 7 - CMSE 381

Prof. Elizabeth Munch

Michigan State University
::
Dept of Computational Mathematics, Science & Engineering

Fri, Sep 13, 2024

Announcements

Last time:

- 3.2 Multiple Linear Regression

Announcements:

- HW #2 Due Sunday!
- Office hours

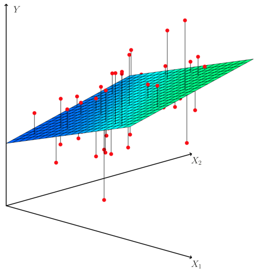
Covered in this lecture

- 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 $\hat{\beta}_0$, $\hat{\beta}_1, \cdots$.
- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \cdots + \hat{\beta}_p x_p$

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

Review: Questions to ask of your model

1. Is at least one of the predictors $X_1, \cdots, X_p$ 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:

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

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

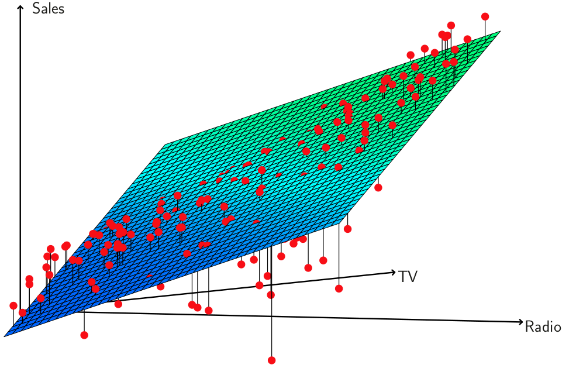
R^2 on Advertising data

- 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?

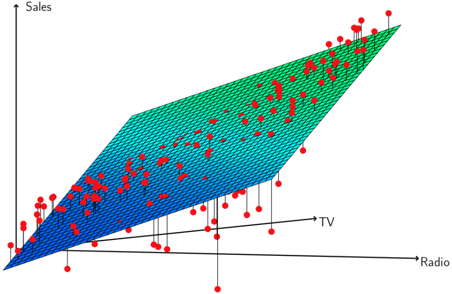
Q4: Making predictions

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

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

estimate for the true population regression plane

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



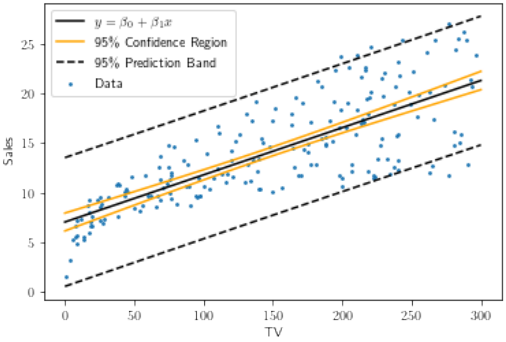
Confidence vs Prediction Model

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**

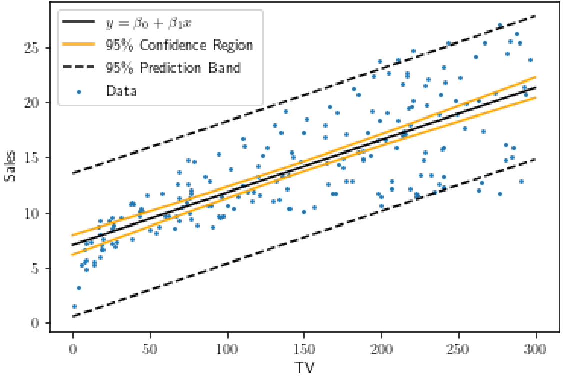
95% CI for sales:
[10,985, 11,528].

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 $X_1, \cdots, X_p$ 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

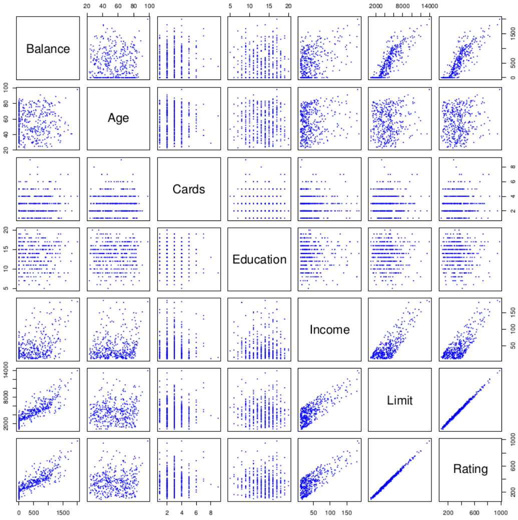
Qualitative Predictors

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: ± 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:

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

Qualitiative Predictor with More than Two Levels

`Region`:

	x_{i1}	x_{i2}
South		
West		
East		

Create spare dummy variables:

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

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

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

More on multiple levels

	Coefficient	Std. error	t-statistic	p-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

Lec #	Date			Reading	HW
1	Mon	8/26	Intro / First day stuff / Python Review Pt 1	1	
2	Wed	8/28	What is statistical learning?	2.1	
	Fri	8/30	Class Cancelled (Dr Munch out of town)		
	Mon	9/2	No class - Labor day		
3	Wed	9/4	Assessing Model Accuracy	2.2.1, 2.2.2	
4	Fri	9/6	Linear Regression	3.1	HW #1 Due Sun 9/8
5	Mon	9/9	More Linear Regression	3.1/3.2	
6	Wed	9/11	Even more linear regression	3.2.2	
7	Fri	9/13	Probably more linear regression	3.3	Hw #2 Due Dun 9/15
8	Mon	9/16	Linear regression coding module		
9	Wed	9/18	Intro to classification, Bayes classifier, KNN classifier	2.2.3	
10	Fri	9/20	Logistic Regression	4.1, 4.2, 4.3.1-3	Hw #3 Due Sun 9/22
11	Mon	9/23	Multiple Logistic Regression / Multinomial Logistic Regression /Project day	4.3.4-5	
	Wed	9/25	***Review***		