Ch 3.3: Even More Linear Regression

Lecture 7 - CMSE 381

Rachel Roca and Prof. Elizabeth Munch

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

Weds, Sep 13, 2023

Announcements

Last time:

- 3.2 Multiple Linear Regression

Announcements:

- Finishing up Linear Regression next time, will be lots of practice
- Office hours Friday (Dr. Munch's office hours canceled)

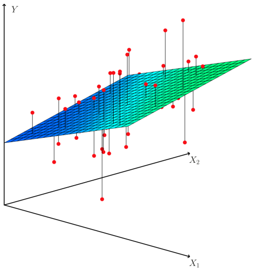
Covered in this lecture

- Qualitative predictors
- Extending the linear model with interaction terms
- Hierarchy principle
- Polynomial regression

Section 1

Review from last time

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*

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?

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:

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

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

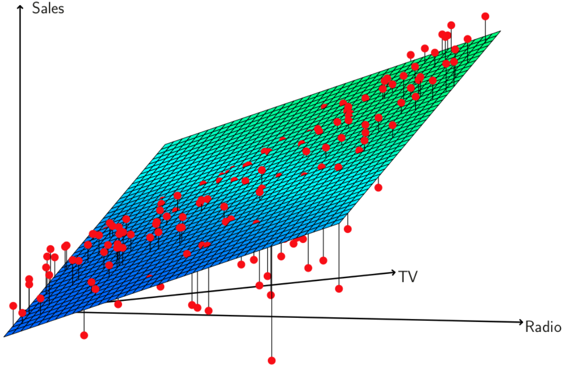
R^2 on Advertising data

- 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



Section 2

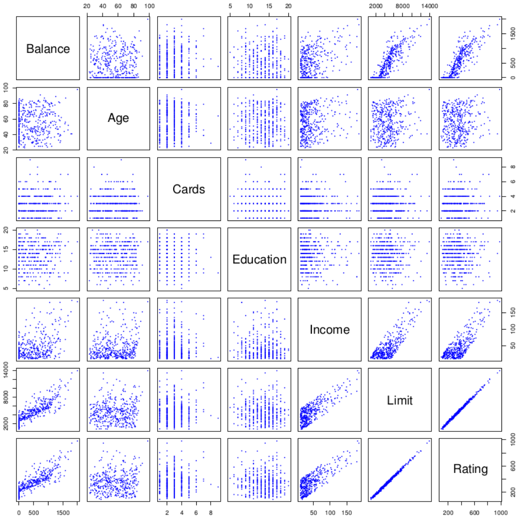
Qualitative Predictors

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:

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

Example

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

Model:

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

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

Model:

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

Qualitiative Predictor with More than Two Levels (possible values)

`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"

Section 3

Extending the linear model

Assumptions so far

Back to our Advertising data set

$$\hat{Y}_{sales} = \beta_0 + \beta_1 \cdot X_{TV} + \beta_2 \cdot X_{radio} + \beta_3 \cdot X_{newspaper}$$

Assumed (implicitly) that the effect on sales by increasing one medium is independent of the others.

What if spending money on radio advertising increases the effectiveness of TV advertising? How do we model it?

Interaction Term

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

$$\begin{aligned} Y_{sales} &= \beta_0 + \beta_1 X_{TV} + \beta_2 X_{radio} + \beta_3 X_{radio} X_{TV} + \varepsilon \\ &= \beta_0 + (\beta_1 + \beta_3 X_{radio}) X_{TV} + \beta_2 X_{radio} + \varepsilon \\ &= \beta_0 + \tilde{\beta}_1 X_{TV} + \beta_2 X_{radio} + \varepsilon \end{aligned}$$

Interaction term

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

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	6.7502	0.248	27.23	< 0.0001
`TV`	0.0191	0.002	12.70	< 0.0001
`radio`	0.0289	0.009	3.24	0.0014
`TV`×`radio`	0.0011	0.000	20.73	< 0.0001

$$\begin{aligned} Y_{sales} &= \beta_0 + \beta_1 X_{TV} + \beta_2 X_{radio} + \beta_3 X_{radio} X_{TV} + \varepsilon \\ &= \beta_0 + (\beta_1 + \beta_3 X_{radio}) X_{TV} + \beta_2 X_{radio} + \varepsilon \end{aligned}$$

Interpretation

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	6.7502	0.248	27.23	< 0.0001
`TV`	0.0191	0.002	12.70	< 0.0001
`radio`	0.0289	0.009	3.24	0.0014
`TV`$\times$`radio`	0.0011	0.000	20.73	< 0.0001

Do the section on “Interaction Terms”

Hierarchy principle

Sometimes p-value for interaction term is very small, but associated main effects are not.

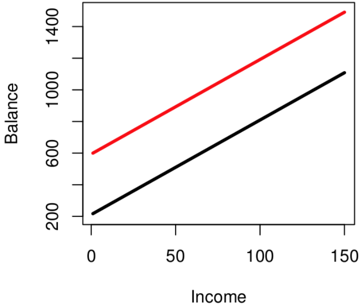
The hierarchy principle:

Interaction term for qualitative variables

Without interaction term

For credit data set:
Predict balance using income (quantitative) and student (qualitative)

$$\begin{aligned}\texttt{balance}_i &\approx \beta_0 + \beta_1 \cdot \texttt{income}_i + \begin{cases}\beta_2 & \text{if student}\\ 0 & \text{if not}\end{cases}\\ &\approx \beta_1 \cdot \texttt{income}_i + \begin{cases}\beta_0 + \beta_2 & \text{if student}\\ \beta_0 & \text{if not}\end{cases}\end{aligned}$$



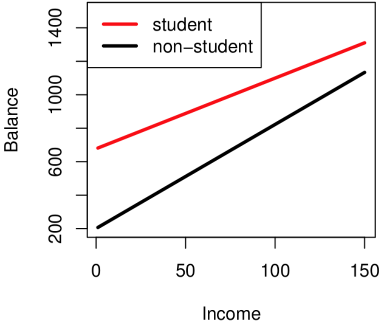
Interaction term for qualitative variables

With interaction term

For credit data set:
Predict balance using income (quantitative) and student (qualitative)

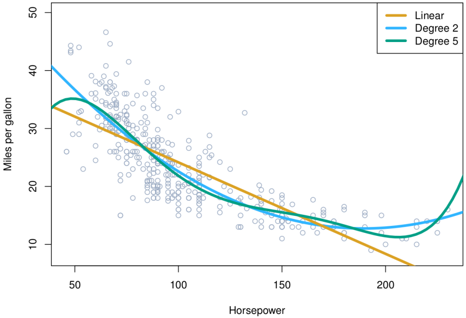
$$\texttt{balance}_i \approx \beta_0 + \beta_1 \cdot \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \cdot \texttt{income}_i & \text{if student} \\ 0 & \text{if not} \end{cases}$$

$$\approx \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \cdot \texttt{income}_i & \text{if student} \\ \beta_0 + \beta_1 \cdot \texttt{income}_i & \text{if not} \end{cases}$$



Nonlinear relationships

$$\texttt{mpg} = \beta_0 + \beta_1 \cdot \texttt{horsepower} + \beta_2 \cdot \texttt{horsepower}^2 + \varepsilon$$



	Coefficient	Std. error	t-statistic	p-value
`Intercept`	56.9001	1.8004	31.6	< 0.0001
`horsepower`	-0.4662	0.0311	-15.0	< 0.0001
$\texttt{horsepower}^2$	0.0012	0.0001	10.1	< 0.0001

Next time

Lec #	Date			Reading
1	Mon	Aug 28	Intro / First day stuff / Python Review Pt 1	1
2	Wed	Aug 30	What is statistical learning?	2.1
	Fri	Sep 1	Assessing Model Accuracy	2.2.1, 2.2.2
3	Mon	Sep 4	No class - Labor day	
4	Wed	Sep 6	Linear Regression	3.1
5	Fri	Sep 8	More Linear Regression	3.1/3.2
6	Mon	Sep 11	Even more linear regression	3.2.2
7	Wed	Sep 13	Probably more linear regression	3.3
8	Fri	Sep 15	Intro to classification, Logisitic Regression	2.2.3, 4.1, 4.2, 4.3
9	Mon	Sep 18	More logistic regression	
10	Wed	Sep 20	Multiple Logistic Regression / Multinomial Logistic Regression	
11	Fri	Sep 22	Overflow/Project day?	
	Mon	Sep 25	***Review***	
	Wed	Sep 27	**Midterm #1**	
	Fri	Sep 29	No class - Dr Munch out of town	
12	Mon	Oct 2	Leave one out CV	5.1.1, 5.1.2
13	Wed	Oct 4	k-fold CV	5.1.3
14	Fri	Oct 6	More k-fold CV,	5.1.4-5
15	Mon	Oct 9	k-fold CV for classification	5.1.5
16	Wed	Oct 11	Resampling methods: Bootstrap	5.2
17	Fri	Oct 13	Subset selection	6.1