More on Multi-Linear Regression

Lecture 5 - CMSE 381

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

January 23, 2024

Announcements

Last time:

- 3.2 Multiple Linear Regression

Announcements:

- Second homework due today
- Office hours

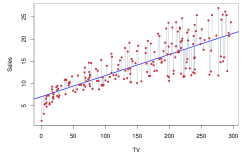
Covered in this lecture

- Multiple linear regression
- Hypothesis test in that case
- Model Selection
- R^2 and RSE
- Confidence intervals and prediction intervals
- Qualitative predictors
- Extending the linear model with interaction terms
- Hierarchy principle
- Polynomial regression

Section 1

Review from last time

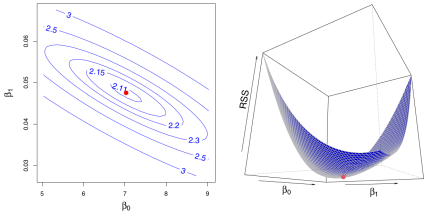
Linear Regression with One Variable



- Predict Y on a single variable X

$$Y \approx \beta_0 + \beta_1 X$$

- Find good guesses for $\hat{\beta}_0$, $\hat{\beta}_1$.
- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- $e_i = y_i - \hat{y}_i$ is the ith residual
- RSS $= \sum_i e_i^2$



- RSS is minimized at *least squares coefficient estimates*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Evaluating the model

- Linear regression is unbiased
- Variance of linear regression estimates:

$$\mathrm{SE}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right]$$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

  where $\sigma^2 = \mathrm{Var}(\varepsilon)$
- Estimate σ: $\hat{\sigma}^2 = \frac{RSS}{n-2}$

- The 95% confidence interval for β_1 approximately takes the form

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$$

- Hypothesis test:
  H_0: $\beta_1 = 0$
  H_a: $\beta_1 \neq 0$
  - Test statistic $t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$

Assessing the accuracy of the model

Residual standard error (RSE):

$$RSE = \sqrt{\frac{1}{n-2}RSS}$$

measure lack of fit;

R squared:

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

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

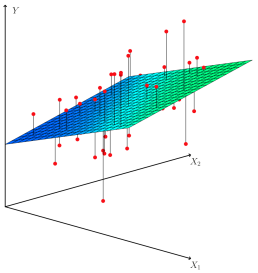
- Proportion in [0,1]; variability in Y that can be explained using X
- 0 means Y not explained by linear regression on X

Least Squares Prediction for Multiple Linear Regression

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

Given estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p$, prediction is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$



Minimize the sum of squares

$$RSS = \sum_i (y_i - \hat{y}_i)^2$$
$$= \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \cdots - \hat{\beta}_p x_p)$$

- Coefficients are closed form but UGLY
- β_j is average effect on Y for one unit increase in X_j if all other X_i stay fixed

Section 2

Ch 3.2.2: Questions to ask of your regression

Question 1

Is at least one of the predictors $X_1, \cdots, X_p$ useful in predicting the response?

Q1: Hypothesis test

H_0 : $\beta_1 = \beta_2 = \cdots = \beta_p = 0$

H_a : At least one β_j is non-zero

F-statistic:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{p,n-p-1}$$

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

The F-statistic for the hypothesis test

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{p,n-p-1}$$

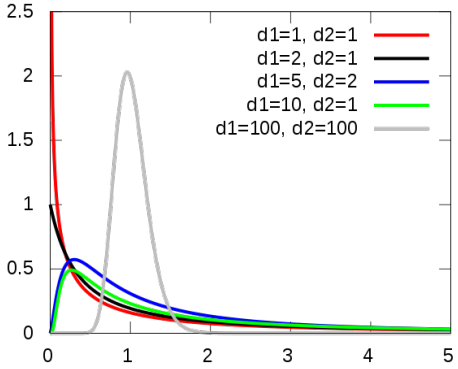


Image from wikipedia, By IkamusumeFan - Own work, CC BY-SA 4.0,

- If linear model assumptions are true

$$E(RSS/(n-p-1)) = \sigma^2$$

- and if H_0 is true

$$E((TSS - RSS)/p) = \sigma^2$$

- So if no relationship between response and predictors, F is close to 1
- If H_a is true,

$$E((TSS - RSS)/p) > \sigma^2$$

so then $F > 1$

Do Q1 section in jupyter notebook

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	2.939	0.3119	9.42	< 0.0001
`TV`	0.046	0.0014	32.81	< 0.0001
`radio`	0.189	0.0086	21.89	< 0.0001
`newspaper`	-0.001	0.0059	-0.18	0.8599

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

- F stat is way bigger than 1, so reject null hypothesis
- How big is big enough? Use the p-value. Not included here but it's basically 0, so we're good

- Can alos point out that how far from 1 required depends on the p, n values.
- If n large, an F-stat a little larger than 1 is still enough evidence against H_0
- However, if n small, might need to be further from 1 to be convinced.

Q2

Do all the predictors help to explain Y, or is only a subset of the predictors useful?

Q2: A first idea

Great, you know at least one variable is important, so which is it?....

- Bad idea plan A: All subsets or best subsets regression
  - Take every possible subset of p variables
  - Figure out least squares fit
  - Choose the best one based on some criterion balancing training error with model size

Why is this a bad idea?

- Why bad idea: $p = 40$ variables, this is checking more than a billion models
- Better idea: Automated approach that searches the subsets for a good one without checking all of them

- Later in the class, we will discuss better options for this (Ch 6)
- Forward selection/ Backward selection to get subsets of variables without checking every possible subset
- Shrinkage methods (Lasso/Ridge)

Q3

How well does the model fit the data?

Assessing the accuracy of the model

Almost the same as before

Residual standard error (RSE):

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

measure lack of fit;

R squared:

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

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

- Proportion in [0,1]; variability in Y that can be explained using X
- 0 means Y not explained by linear regression on X

R^2 on Advertising data

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

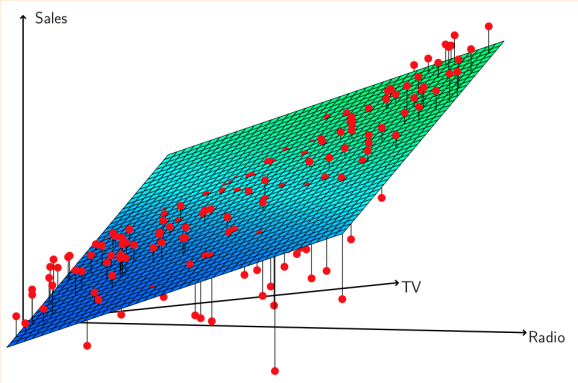
- R^2 always increases with more variables, even if those variables are weakly associated to the response
- Adding another variable decreases RSS on training
- Small increase in R^2 if we include newspaper to other two, so can safely ignore
- Large increase going from just TV to TV and Radio, so likely want to keep that

RSE on Advertising Data

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

- Similar argument about big jump from TV to TV and radio, but small difference after that
- Even though RSS went down, RSE goes up at the end because increase in p made a difference in the fraction $RSE = \sqrt{\frac{RSS}{n-p-1}}$

If all else fails, look at the data



- Model overestimates sales for cases where either spent only on TV, or only spent on radio
- Underestimates sales where budget was split between the two
- Implies there are some interaction effects between the two
- Later we'll extend the model to allow for this interaction

Q4

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

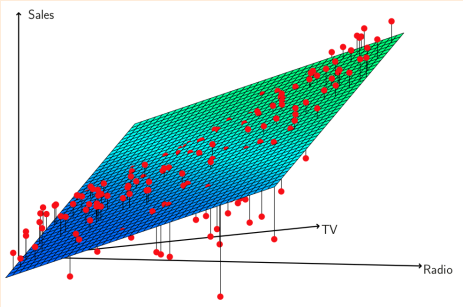
Q4: Making predictions

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 intervals and Prediction Intervals

Confidence Interval

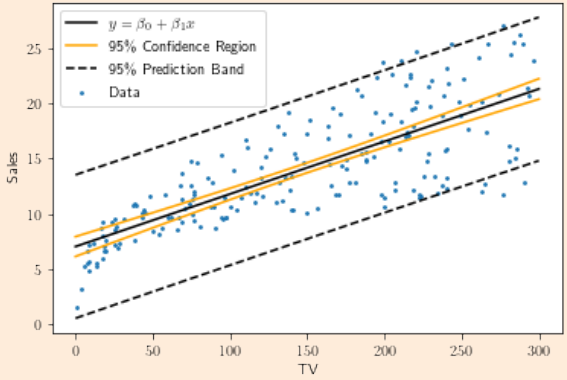
The range likely to contain the population parameter (mean, standard deviation) of interest.

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.

Comparing the two

- Example showing the 1-d regression case
- Fixing TV = \$150, confindence interval for sales is tiny
- High probability the correct model takes a value in there
- prediction interval is large
- Need wider band to be confident that a new observation (with noise) is contained in there



Section 3

Qualitative Predictors

Reminder: Qualitative vs Quantitative predictors

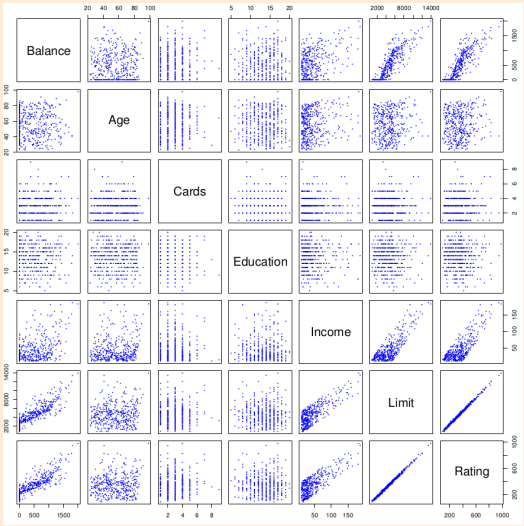
Quantitative:

- Anything with values in either integers or real numbers
- Temperature
- Year (even though integers)
- Weight
- Number of cyllendars in a car

Qualitative/Categorical:

- Anything from an unordered set
- Country of origin
- Brand names
- Anything yes/no (married, has a cat)

New data set! Credit card balance



- `own`: house ownership *This is referenced in the text, but isn't in any of the online versions of the data set that I can find*
- `student`: student status
- `status`: marital status
- `region`: East, West, or South *Also referenced in the book, but not in the online data sets*

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 called dummy 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:

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

- β_0 is avg credit card balance among those who are not a student
- $\beta_0 + \beta_1$ is the avg for those are
- β_1 is avg diff in balance

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

- Avg CC balance for non-students \$480.36,
- Students have \$396.46 additional debt, so \$876.83

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:

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

Note that the predictions will be the same

Qualitiative Predictor with More than Two Levels

`Region`:

	x_{i1}	x_{i2}
South	*1*	*0*
West	*0*	*1*
East	*0*	*0*

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

- Always one less variable than levels
- Level with no dummy variable is called the baseline

	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

- Balance for the baseline (EAST) is \$531
- South has \$18.69 less than east
- West has \$12.50 less than east
- p-values for dummy variables are large, so can't reject null hypothesis
- No statistical evidence that there's a diff in CC balance in different regions

Section 4

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

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

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

- Draw a matrix over here where we've made a new column $X_{radio} X_{TV}$

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

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

- P-value for interaction term is low, strong evidance for $H_a : \beta_3 \neq 0$.
- R^2 is 96.8% for this, vs 89.7% for model without interaction

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×radio`	0.0011	0.000	20.73	< 0.0001

- Increase in \$1K of TV is associated to increased sales of

$$(\beta_1 + \beta_3 X_{radio}) \cdot 1000 = 19 + 1.1 \cdot X_{radio}$$

units

- Increase in radio advertising of \$1K will be associated with an increase sales of

$$(\beta_2 + \beta_3 X_{TV}) \cdot 1000 = 29 + 1.1 \cdot X_{TV}$$

Hierarchy principle

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

The hierarchy principle:

- If we include the interaction term, we should include the main effects no matter what.
-

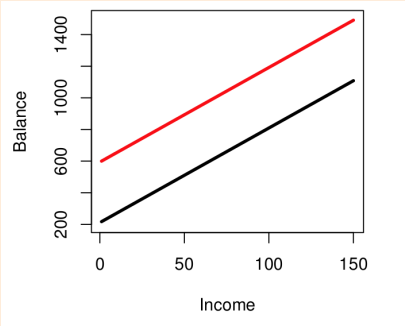
Interaction term for qualitative variables

Without interaction term

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

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

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



Fitting two parallel lines to the different subsets of data

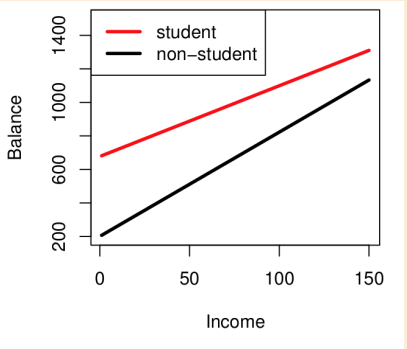
Interaction term for qualitative variables

With interaction term

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

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

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

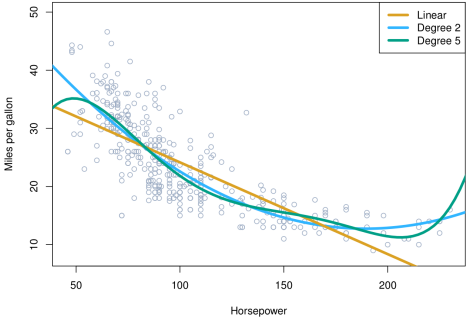


Fitting two different lines to the different subsets of data

Nonlinear relationships

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

- Quadratic shape of data suggests above model
- Still linear because we're doing multi-linear regression with $X_1 = \textit{horsepower}$ and $X_2 = \textit{horsepower}^2$
- Could add more terms, but risk of overfitting
- Polynomial regression, more in ch 7



	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
horsepower^2	0.0012	0.0001	10.1	< 0.0001