Ch 3.1-2: (Multi)-Linear Regression

Lecture 4 - CMSE 381

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

Announcements

Last time:

- Started 3.1 - Single linear regression

Announcements:

- Office Hours: Monday-Thursday
- Homework #1 grades and feedback posted
- Homework #2 Due Wed, Jan 24

Covered in this lecture

- hypothesis test, and p-value for coefficient estimates
- Residual standard error (RSE)
- R squared
- Setup for multiple linear regression

Section 1

Last time

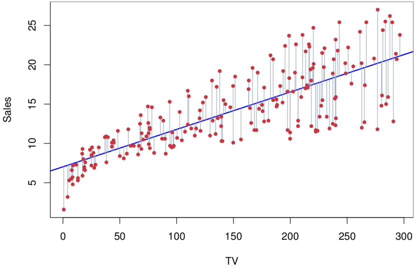
Setup

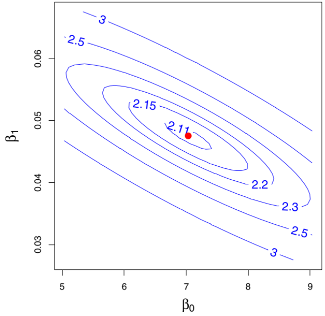
- Predict Y on a single predictor variable X

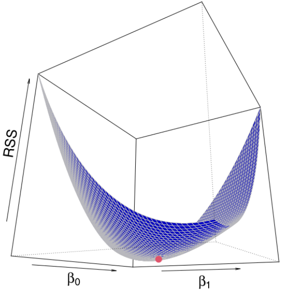
$$Y \approx \beta_0 + \beta_1 X$$

- "$\approx$" .... "is approximately modeled as"

- Given $(x_1, y_1), \cdots, (x_n, y_n)$
- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be prediction for Y on ith value of X.
- $e_i = y_i - \hat{y}_i$ is the ith residual







Residual sum of squares RSS is

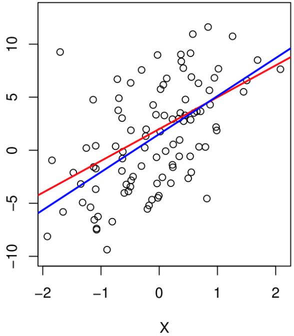
$$RSS = e_1^2 + \cdots + e_n^2$$
$$= \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

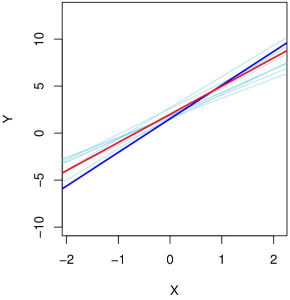
Least squares criterion

Find β_0 and β_1 that minimize the RSS.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Linear regression is unbiased





- 100 data points drawn from $Y = 2 + 3X + \varepsilon$
- ε drawn from normal distribution with mean 0
- Red line is true relationship, blue is least squares estimate
- Repeat this 10 times and plot all the found lines (in variations of blue)
- The resulting models are slightly different but are all around the red true relationship

Section 2

Continue on evaluating models

Variance of linear regression estimates

- 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 $\sigma^2 = \mathrm{Var}(\varepsilon)$

- Residual standard error is an estimate of σ

$$RSE = \sqrt{RSS/(n-2)}$$

Confidence Interval

The 95% confidence interval for β_1 approximately takes the form

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

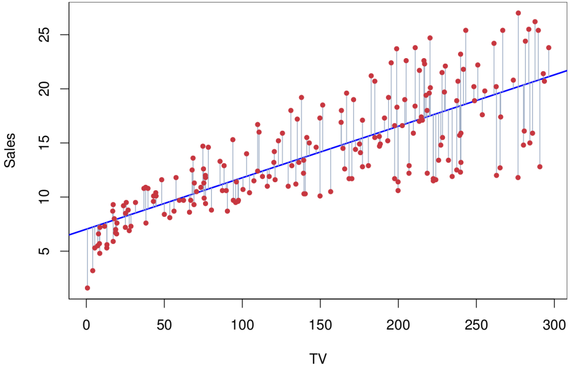
Same form works for β_0

Interpretation:
There is approximately a 95% chance that the interval

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

will contain β_1 where we repeatedly approximate $\hat{\beta}_1$ using repeated samples.

CI in Advertising data



For the advertising data set, the 95% CIs are:

- β_1 :: $[0.042, 0.053]$
- β_0 :: $[6.130, 7.935]$

Hypothesis testing

H_0: There is no relationship between X and Y *(null hypothesis)*

H_1: There is some relationship between X and Y *(alternative hypothesis)*

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

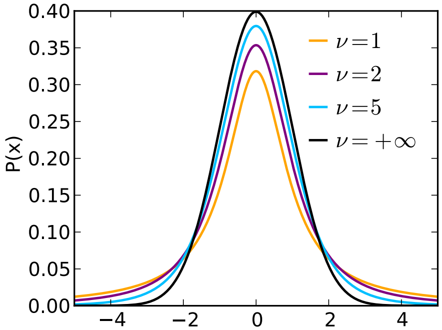
since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \varepsilon$, and thus X is not associated with Y

Test statistic and p-value

Test statistic:

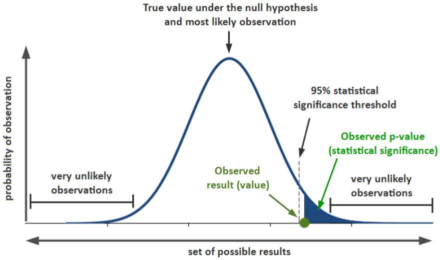
$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

t-distribution with $n-2$ degrees of freedom



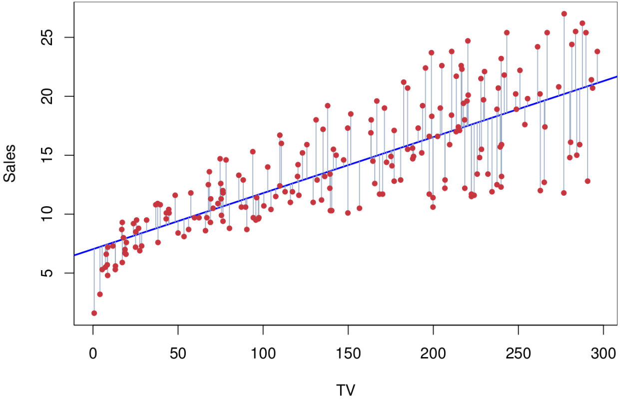
A small p-value indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance, in the absence of any real association between the predictor and the response.

Draw me:



Advertising example

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	7.0325	0.4578	15.36	< 0.0001
`TV`	0.0475	0.0027	17.67	< 0.0001



Assessing the accuracy of the module: RSE

Quantify the extent to which the model fits the data

Residual standard error (RSE):

$$
\begin{aligned}
RSE &= \sqrt{\frac{1}{n-2} RSS} \\
&= \sqrt{\frac{1}{n-2} \sum_i (y_i - \hat{y}_i)^2}
\end{aligned}
$$

- estimate of the standard deviation of ε
- Avg amount that the response will deviate from the true regression line
- avg amount response will deviate from the true regression line

Assessing the accuracy of the module: R^2

R squared:

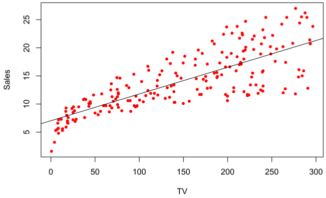
$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

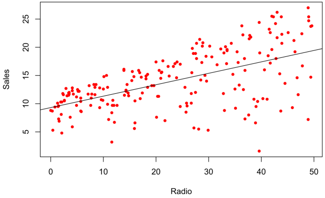
where total sum of squares is

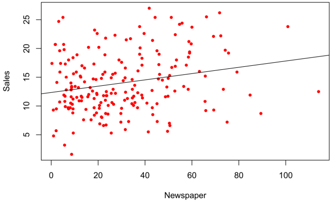
$$TSS = \sum_i (y_i - \overline{y})^2$$

- TSS is total variance in teh response Y, variability before regression
- RSS amount of variability after the regression
- R^2 is proportion of variability in Y that can be explained using X
- Close to 1, large proportion of varaiability is explained by regression
- Close to 0, regression does not explain much of the variability in the response

Advertising example







$R^2 = 0.61$ $\qquad$ $R^2 = 0.33$ $\qquad$ $R^2 = 0.05$

Coding group work

Run the section titled "Assessing Coefficient Estimate Accuracy"

Point out that the homework uses the code slightly differently. statsmodels.formula.api vs statsmodels.api. You can use whatever you want on the homework.

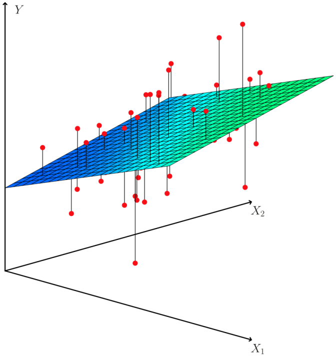
Section 3

Multiple Linear Regression

Setup

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

- β_j is avg affect on Y of one unit increase in X_j



Estimation and Prediction

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

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

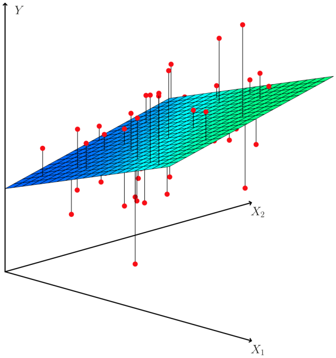
Minimize the sum of squares

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

Coefficients are closed form but UGLY
We won't write them down, your favorite code can do this for us

Advertising data set example

$$\texttt{Sales} = \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{radio} + \beta_3 \cdot \texttt{newspaper}$$



	Coefficient
Intercept	2.939
TV	0.046
radio	0.189
newspaper	−0.001

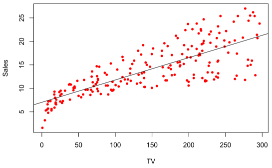
Interpretation of coefficients

$$\texttt{Sales} = \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{radio} + \beta_3 \cdot \texttt{newspaper}$$

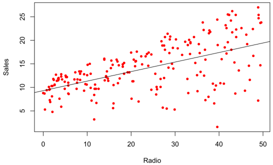
	Coefficient
Intercept	2.939
TV	0.046
radio	0.189
newspaper	−0.001

- Fixing TV and newspaper spending; Spending $1K more on radio results in 189 units additional sales
- What's going on with newspaper? This says no relationship between newspaper and sales

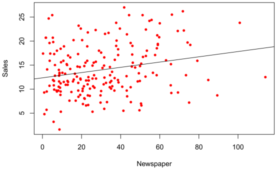
Single regression vs multi-regression



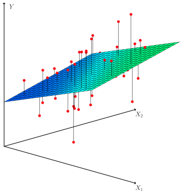
	Coefficient
Intercept	7.0325
TV	0.0475



	Coefficient
Intercept	9.312
radio	0.203



	Coefficient
Intercept	12.351
newspaper	0.055



	Coefficient
Intercept	2.939
TV	0.046
radio	0.189
newspaper	−0.001

Correlation matrix

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

- Correlation between radio and newspaper high (0.35)
- Markets with lots of radio also have lots of newspaper ads
- In single reg, newspaper gets credit for radio

Coding group work

Run the section titled "Multiple Linear Regression"