CSCI 3022

intro to data science with probability & statistics

November 26, 2018

Inference & Model Selection in Multiple Linear Regression

Ankaine ☺

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Stuff & Things

1. **Homework 6** is due Friday. Final homework! :D
2. **Final Exam** (Dan's section): December 18, Tuesday, 7:30 PM to 10 PM.
3. **Practicum** posted tonight. Due **Wednesday, 12/12, 11:55 PM.** *No late submissions accepted.*

Practicum Rules

Arkaive :

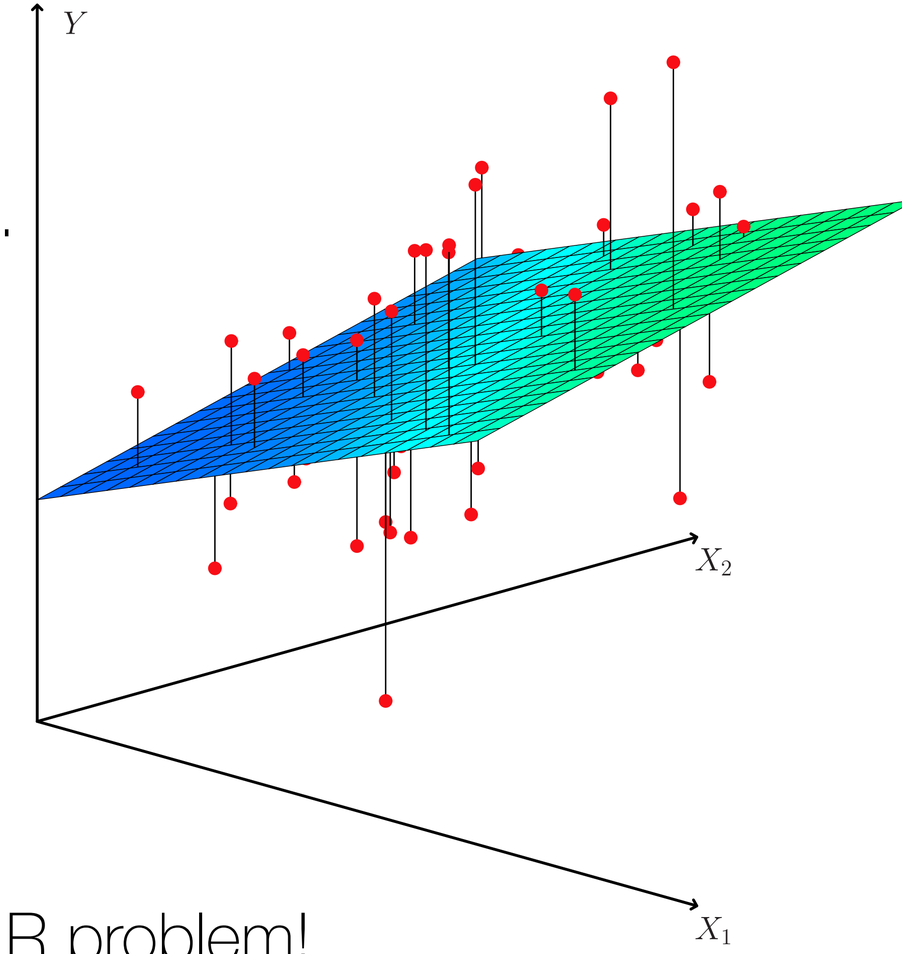
1. All work, code and analysis must be **your own**.
2. You may use your course notes, posted lecture slides, textbooks, in-class notebooks, and homework solutions as resources. You may also search online for answers to general knowledge questions like the form of a probability distribution function or how to perform a particular operation in Python/Pandas.
3. You may **not** post to message boards or other online resources asking for help.
4. **You may not collaborate with classmates or anyone else.**
5. This is meant to be like a coding portion of your final exam. So, we will be much less helpful than we typically are with homework. For example, we will not check answers, help debug your code, and so on.
6. If you have a question, send me/Tony a **private** Piazza message. If we decide that it is appropriate for the entire class, then we will add it to the **Practicum Q&A (@314)**.
7. If something is left open-ended, it is because we want to see how you approach the kinds of problems you will encounter in the wild, where it will not always be clear what sort of tests/methods should be applied. Feel free to ask clarifying questions though.

Last time on CSCI 3022:

- Multiple Linear Regression assumes that the response y may be affected by multiple features.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

- Instead of fitting a line to the data, MLR fits a plane.
- What did we learn about MLR vs SLR?



- Note: we can cast *polynomial regression* as an MLR problem!

Recap: advertising budgets

SLR

```
SLR for tv vs sales
----------------------
intercept = 7.0326
slope = 0.0475
p-value = 1.4673897001945922e-42


SLR for radio vs sales
----------------------
intercept = 9.3116
slope = 0.2025
p-value = 4.354966001766913e-19


SLR for news vs sales
----------------------
intercept = 12.3514
slope = 0.0547
p-value = 0.0011481958688882112
```

MLR

$$\texttt{sales} = 2.94 + 0.046 \times \texttt{TV} + 0.189 \times \texttt{radio} - 0.001 \times \texttt{news}$$

Under SLR, each feature shows a significant slope.
Under MLR, the coefficient for newspapers disappears.

Covariance and Correlation of Features

- One way to discover this relationship between features is to do a **correlation analysis**. We want to know, if the value of one feature goes up is it likely that the other feature will go up as well? Similarly, we might find that if one feature goes up is it likely that the other feature will go down?
- **Def**: Let X and Y be random variables. The covariance between X and Y is given by

$$Cov(X,Y) = E\Big[(X - E[X])(Y - E[Y])\Big]$$

Let $Y=X$ $\quad Cov(X,X) = E\Big[(X-E[X])(X-E[X])\Big] = E\Big[(X-E[X])^2\Big] = Var(X)$

- **Def**: The correlation coefficient $\rho(X,Y)$ is a measure between -1 and 1, given by

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Estimating Covariance and Correlation

- We can estimate these relationships from the data using formulas analogous to the sample variance.
- **Def**: The sample covariance is given by

$$S_{xy} = \frac{1}{n-2} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

- **Def**: The sample correlation coefficient is then given by

$$\hat{\rho}_{xy} = \frac{S_{xy}}{\sqrt{S_x S_y}}$$

sample variances (calculated as before)

Advertising Budget Example

- Let's compute the pairwise correlation coefficients for the TV, radio, and newspaper spending features in the advertising data.

```
In [40]:  1 dfAd[["tv", "radio", "news"]].corr()
```

Out[40]:

	tv	radio	news
tv	1.000000	0.054809	0.056648
radio	0.054809	1.000000	0.354104
news	0.056648	0.354104	1.000000

- **Question**: What do you notice? radio and news are correlated!

Recap: advertising budgets

SLR

```
SLR for tv vs sales
----------------------
intercept = 7.0326
slope = 0.0475
p-value = 1.4673897001945922e-42


SLR for radio vs sales
----------------------
intercept = 9.3116
slope = 0.2025
p-value = 4.354966001766913e-19


SLR for news vs sales
----------------------
intercept = 12.3514
slope = 0.0547
p-value = 0.0011481958688882112
```

MLR

$$\texttt{sales} = 2.94 + 0.046 \times \texttt{TV} + 0.189 \times \texttt{radio} - 0.001 \times \texttt{news}$$

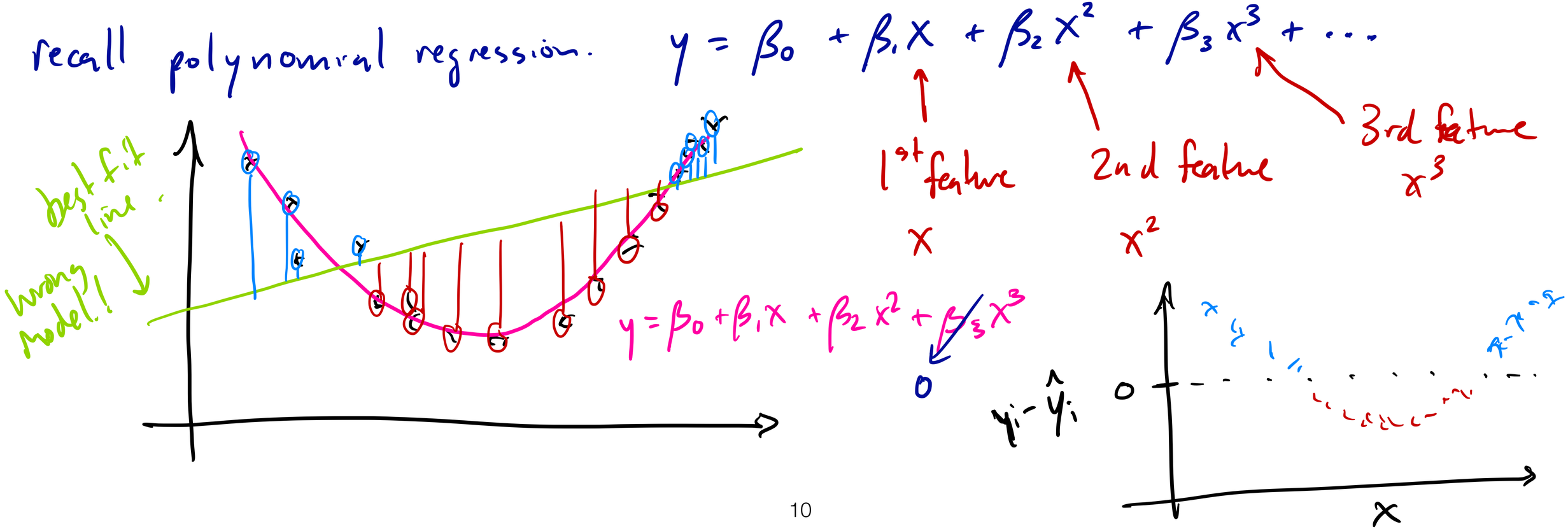
Under SLR, each feature shows a significant slope.
Under MLR, the coefficient for newspapers disappears.

This is because *news* is a surrogate for *radio*, which we learned from the correlation matrix.

	tv	radio	news
tv	1.000000	0.054809	0.056648
radio	0.054809	1.000000	0.354104
news	0.056648	0.354104	1.000000

Polynomial regression

- For single-feature data, we can fit a polynomial regression model by casting it as a multiple linear regression where the additional features are powers of the original single-feature, *x*.



Using Residual Plots in Polynomial Reg.

- Recall that the assumed nature of our true model is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dots + \beta_p x^p + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

If true model is $y = \beta_0 + \beta_1 x + \varepsilon$

and our model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

then $r = y - \hat{y} = \varepsilon \sim N(0, \sigma^2)$

If true model is $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

and our model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$r = y - \hat{y} \sim N(\beta_2 x^2, \sigma^2)$

See last problem on previous notebook

If I plot the residuals $r_i = y_i - \hat{y}_i$, these should be normally distributed with no dependence on x, when my model is correct.

Inference in Multiple Linear Regression

- Questions we would like to answer:
  1. Is at least one of the features useful in predicting the response?
  2. Do all of the features help to explain the response, or is it just a subset?
  3. How well does the model fit the data?

Hypothesis Testing for MLR

- Recall our question from ~~last~~ this time:

Is there a relationship between the response and predictors?

- In the simple linear regression setting, we can simply check whether $\beta_1 = 0$.
- In the MLR setting, with *p* features (aka predictors) we n eed to ask whether *all* of the coefficients are zero:
  - H_0 : $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_p = 0$
  - H_1 : At least one β is non-zero. $\Rightarrow$ $\beta_j \neq 0$ for at least one value of $j \neq 0$

H_1 is not $\beta_1 \neq 0$, $\beta_2 \neq 0$, $\beta_3 \neq 0$ ...

Is at Least One Feature Important?

- We test the hypothesis via the F-statistic.

$$F = \frac{(SST - SSE)}{df_{SST} - df_{SSE}} \Big/ \frac{SSE}{df_{SSE}} = \frac{(SST - SSE)/p}{SSE/(n-p-1)} = F$$

- Recall:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip})\right)^2$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad df_{SST} = n - 1$$

$$df_{SSE} = n - (p+1) = n - p - 1$$

Is at Least One Feature Important?

- We test the hypothesis via the F-statistic.

$$F = \frac{(SST - SSE)/p}{SSE/(n-p-1)} \qquad SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2$$

- Suppose H_0 were true. What would F be?

F around 1

- Suppose that H_1 were true. What would F be?

F > 1

The F-statistic

- We test the hypothesis via the F-statistic.

two different d.of. parameters.

$$\tilde{F} = \frac{(SST - SSE)/p}{SSE/(n-p-1)}$$

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2$$

- F distribution will give us a critical value so that we can do a p-value or rejection region test.

→ Always a one-tailed test. $F \overset{?}{\geq} F_{critical}$

compare to α like we normally would

$1 - \text{scipy.stats.f.cdf}(\tilde{F}, p, n-p-1)$ ← $Pr(\tilde{F} \geq F_{p,n-p-1})$