CSCI 3022

intro to data science with probability & statistics

Lecture 25 (TWENTY FIVE?!?)
April 16, 2018

Inference & Model Selection in Multiple Linear Regression

"Dreams come true!"

- you, in ref. to today's ~~exciting~~ astounding material.

Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Dan Larremore

Stuff & Things

- HW6 due Friday.
- No more homework! Instead: practicum, posted ~~on~~ Tonight.
  - Real data & real questions. Combining EDA, stats, pandas, and all your favorite tricks. A little Monte Carlo simulation. Other fun things to practice data science, "where the rubber meets the code."
  - 3 problems.
  - Due Weds before finals week, 5/2.
  - No collaborating. Sry. Show me what you've learned! 🎉😎🤩 (Ask in OH!)

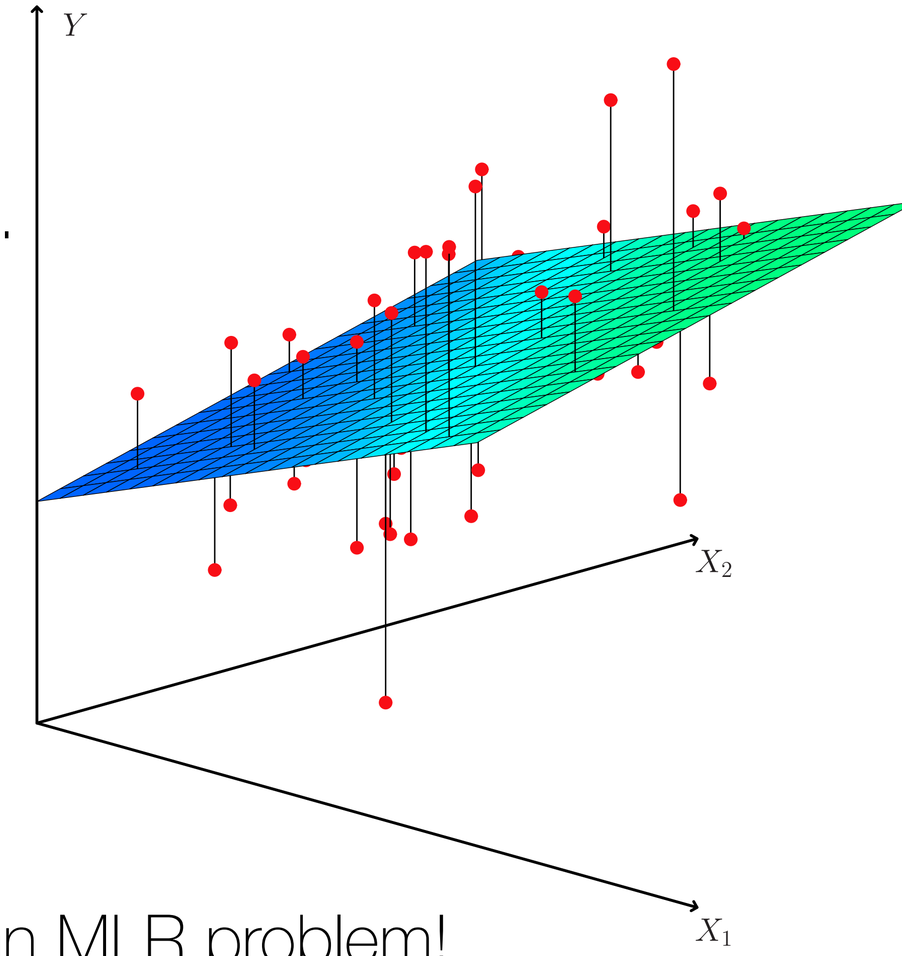
No late days.

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?



- ~~Recall~~ that we can cast *polynomial regression* as an MLR problem!

Polynomial regression

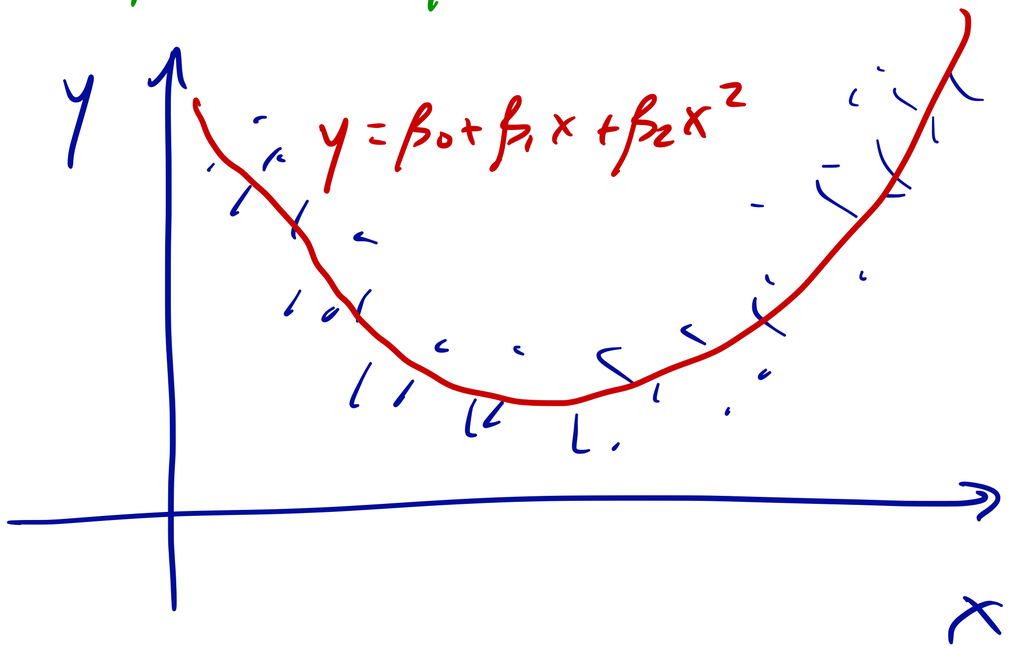
- For single-feature data, we can fit a polynomial regression model by casting it as a multiple linear regression where the additional features are powers of the original single-feature, *x*.

recall polynomial: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots$

first feature $x_1 = x$

second feature $x_2 = x^2$ ←

third feature $x_3 = x^3$



Using Residual Plots in Polynomial Reg.

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

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \cdots + \beta_p x_1^p + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

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

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

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

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

$\Rightarrow$ If I plot the residual (x_i, r_i) should be normally distr. around missing feature.

See last notebook Prob. #3

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.

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

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 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 is not 0, so $\beta_j \neq 0$ for at least one j

this 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.

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

$df_{SST}: n-1$

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

$df_{SST} - df_{SSE} = p$

- Recall:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad df: n-1$$

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

$$= \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi})\right)^2 \quad df: n-(p+1) = n-p-1$$

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 nearly 1

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

F > 1

The F-statistic

- We test the hypothesis via the F-statistic.

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

two diff d.o.f. parameters.

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

- F distribution will give us a critical value so we can do a p-value test!

Is $F \geq F_{critical}$? Always one tailed.

compare this to α

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

Is a Subset of Features Important?

- **Full Model**: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ (p=4 features in full model)
- **Reduced Model**: $y = \beta_0 + \beta_2 x_2 + \beta_4 x_4$ (k=2 features in reduced model)
  - **Question**: Are the missing features important, or are we OK going with the reduced model?
  - **Partial F-Test**: $H_0 : \beta_1 = \beta_3 = 0$
  - Since the features in the reduced model are also in the full model, we expect the full model to perform at least as well as the reduced model.
  - **Strategy**: Fit the Full and Reduced models. Determine if the difference in performance is real or due to just chance.

dropped p-k features

df_{full}: $n-(p+1)$

df_{red}: $n-(k+1)$

$n-(p+1)-[n-(k+1)]$

$n-p-1-n+k+1$

$k-p$ $\quad$ $p-k$

Is a Subset of Features Important?

p is # features in full model

k = features in reduced model

- SSE_{full} = variation unexplained by the full model

- SSE_{red} = variation unexplained by the reduced model

Intuitively, if SSE_{full} is much smaller than SSE_{red}, the full model fits the data much better than the reduced model. The appropriate test statistic should depend on the difference $SSE_{red} - SSE_{full}$ in unexplained variation.

- Test Statistic:

$$F = \frac{(SSE_{\text{red}} - SSE_{\text{full}})/(p-k)}{SSE_{\text{full}}/(n-p-1)} \sim F_{p-k,n-p-1}$$

- Rejection Region:

$$F \geq F_{\alpha,p-k,n-p-1}$$

stats.f.ppf(signif, dof1, dof2)

http://homepage.divms.uiowa.edu/~mbognar/applets/f.html

F… why even?

- Why compute the p-value for F-statistic when instead, we already have p-values for each of the covariates?

- Doing so would not be testing one hypothesis, but rather *p* hypotheses!

- At α=0.05, how many *p* values do we expect to be significant if the null hypothesis is, in fact, true?

p·0.05, so if 100 features, 100·0.05=5

```
In [27]: 1 model.summary()
```

Out[27]:

OLS Regression Results

Dep. Variable:	sales	**R-squared:**	0.897
Model:	OLS	**Adj. R-squared:**	0.896
Method:	Least Squares	**F-statistic:**	570.3
Date:	Tue, 28 Nov 2017	**Prob (F-statistic):**	1.58e-96
Time:	20:28:02	**Log-Likelihood:**	-386.18
No. Observations:	200	**AIC:**	780.4
Df Residuals:	196	**BIC:**	793.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.9389	0.312	9.422	0.000	2.324	3.554
tv	0.0458	0.001	32.809	0.000	0.043	0.049
radio	0.1885	0.009	21.893	0.000	0.172	0.206
news	-0.0010	0.006	-0.177	0.860	-0.013	0.011

The road to R^2 for MLR

- Just as with simple regression, the error sum of squares is:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad \hat{\sigma}^2 = \frac{SSE}{n-(p+1)} = \frac{SSE}{n-p-1}$$

You may see SSE written as RSS: "residual sum of squares"

- It is again interpreted as a measure of how much variation in the observed y values is not explained by (not attributed to) the model relationship.
- The number of df associated with SSE is n–(p+1) because p+1 df are lost in estimating the p+1 β coefficients.

The road to R^2

- Just as before, the **total sum of squares** is:

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

- And the **sum of squared errors** is:

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

- Then the coefficient of multiple determination R^2 is:

$$R^2 = \frac{SSR}{SST} = \boxed{1 - \frac{SSE}{SST} = R^2}$$

- It is interpreted in the same way as before. (Do you remember?)

$\frac{SSE}{SST}$ ~ how much variance is left unexplained after model fit.

Hacking R^2

Unfortunately, there is a problem with R^2: Its value can be inflated by adding lots of predictors into the model even if most of these predictors are frivolous!

Hacking R^2

- For example, suppose y is the sale price of a house. Then:
- Sensible predictors include
  x_1 = the interior size of the house,
  x_2 = the size of the lot on which the house sits,
  x_3 = the number of bedrooms,
  x_4 = the number of bathrooms, and
  x_5 = the house's age.
- But now suppose we add in
  x_6 = the diameter of the doorknob on the coat closet,
  x_7 = the thickness of the cutting board in the kitchen,
  x_8 = the thickness of the patio slab.

Adjusted R^2

- The objective in multiple regression is not simply to explain most of the observed y variation, but to do so using a model with relatively few predictors that are easily interpreted.
- It is thus desirable to adjust R^2 to take account of the size of the model:

$$R^2 = 1 - \frac{SSE}{SST}$$

$$R^2_a = 1 - \frac{SSE/df_{SSE}}{SST/df_{SST}} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

↑ adjusted

Adjusted R^2

- The objective in multiple regression is not simply to explain most of the observed y variation, but to do so using a model with relatively few predictors that are easily interpreted.
- It is thus desirable to adjust R^2 to take account of the size of the model:

$$R_a^2 = 1 - \frac{SSE/df_{SSE}}{SST/df_{SST}} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

Adjusted R^2

```
In [27]:  1 model.summary()
```

Out[27]:

OLS Regression Results

Dep. Variable:	sales	**R-squared:**	0.897
Model:	OLS	**Adj. R-squared:**	0.896
Method:	Least Squares	**F-statistic:**	570.3
Date:	Tue, 28 Nov 2017	**Prob (F-statistic):**	1.58e-96
Time:	20:28:02	**Log-Likelihood:**	-386.18
No. Observations:	200	**AIC:**	780.4
Df Residuals:	196	**BIC:**	793.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.9389	0.312	9.422	0.000	2.324	3.554
tv	0.0458	0.001	32.809	0.000	0.043	0.049
radio	0.1885	0.009	21.893	0.000	0.172	0.206
news	-0.0010	0.006	-0.177	0.860	-0.013	0.011

Deciding on important variables

- Suppose that we have 100 data points (n=100), but we have 200 different features (p=200). How can we learn which features are important and which are not?
- **Some options**:
  - Try all the possible combinations of features in models to see which gives the best fit.

Bad idea! Reason 2^p different models.

$p=3 \rightarrow 8$ models

$p=30 \rightarrow 1,073,741,824$ models.