Inference and Model Selection in Multiple Linear Regression

Administrivia

- **Homework 6** due Friday.
- **Practicum** posted on Monday. Due Wednesday December 13th

Previously on CSCI 3022

Given data $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$ for $i = 1, 2, \ldots, n$ fit a MLR model of the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i \qquad \text{where} \qquad \epsilon_i \sim N(0, \sigma^2)$$

Estimates of the parameters are estimated by minimizing

$$SSE = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i + \cdots \beta_p x_p)\right]^2$$

The covariance and correlation coefficient for random variables X and Y are given by

$$\text{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])] \qquad \text{and} \qquad \rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Recap of Advertising Budget Example

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

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

- SLR: Each advertising medium shows a significant slope
- MLR: The coefficient for newspaper ads disappears

Recap of Advertising Budget Example

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

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

- SLR: Each advertising medium shows a significant slope
- MLR: The coefficient for newspaper ads disappears
- This is because in the SLR news is a surrogate for radio, which we learned by looking at pairwise correlation coefficients

	tv	radio	news
tv	1.000000	0.054809	0.056648
radio	0.054809	1.000000	$\boxed{0.354104}$
news	0.056648	0.354104	1.000000

Inference in Multiple Linear Regression

Questions we would like to answer:

- Is at least one of the features useful in predicting the response?
- Do all of the features help to explain the response, or is it just a subset?
- How well does the model fit the data?

Is at Least One Feature Important?

- In the SLR setting, we can do a hypothesis test to determine if $\beta_1 = 0$
- In the MLR setting with p features, we need to check whether ALL coefficients are zero

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

$H_1:$ $\beta_k \neq 0$ FOR AT LEAST ONE VAL OF k

Is at Least One Feature Important?

The **F-Test**:

- We test the hypothesis via the F-statistic: $F = \dfrac{(SST - SSE)/p}{SSE/(n-p-1)}$

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

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

Is at Least One Feature Important?

The **F-Test**:

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

- Suppose H_0 were true. What would F be?

$\beta_1 = \beta_2 = \cdots = \beta_p = 0 \qquad F \approx 1$

- Suppose H_1 were true. What would F be?

BETTER EXPLAIN DATA $\Rightarrow$ SSE $\downarrow$ $\Rightarrow$

$(SST - SSE) \uparrow \Rightarrow F \uparrow$

Is at Least One Feature Important?

α significance level

The **F-Test:**

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

$$\underbrace{F}_{\text{test statistic}} \sim \underbrace{F}_{\text{F-dist}}{}_{p,\, n-p-1}$$

$p,\ n-p-1$: degrees of freedom.

If $F > F_{\alpha, p, n-p-1}$ then reject H_0 and conclude at least one feature is imp.

p-value = 1 - stats.f.cdf(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.

Is a Subset of Features Important?

K = # FEATURES in REDUCED model

- $SSE_{\text{full}} =$ variation unexplained by the full model
- $SSE_{\text{red}} =$ 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.

$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$

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

← df $n-(p+1)$

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

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

$$F = \frac{(SSE_{red} - SSE_{full}) / (df_{RED} - df_{full})}{SSE_{full} / df_{full}}$$

$$\left.\begin{aligned} df_{RED} &= n-k-1 \\ df_{FULL} &= n-p-1 \end{aligned}\right\}$$

$$df_{red} - df_{FULL} = (\cancel{n}-k-\cancel{1}) - (\cancel{n}-p-\cancel{1})$$

$$= (p-k)$$

Why Use the F-Tests?

- Why compute the p-value for the F-statistic when we could compute p-values for each of the feature slopes?

- If we do this, we're testing p different hypotheses instead of a single hypothesis

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

```
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

Why Use the F-Tests?

- Why compute the p-value for the F-statistic when we could compute p-values for each of the feature slopes?
- If we do this, we're testing p different hypotheses instead of a single hypothesis
- At $\alpha = 0.05$, how many p-values do we expect to be significant if the null hypothesis is, in fact true?
- This is called

  The Problem of Multiple Comparisons

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

What about Goodness-of-Fit?

- Like in SLR, the MLR sum of squared errors is:

$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

- Like in SLR, the MRL total some of squares is:

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

- Then the coefficient of determination is:

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

- It is interpreted as the fraction of variation that **IS** explained by the model

What about Goodness-of-Fit?

Problem: The standard R^2 value you can be artificially inflated by adding lots and lots of frivolous features.

Example: Suppose that y represents the sale price of a house. Reasonable features associated with sale price might be:

- 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 in the house
- x_4: the number of bathrooms in the house
- x_5: the age of the house

But suppose we also add:

- 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

What about Goodness-of-Fit?

- The objective of multiple linear regression is not simply to explain the most variation in the data, but to do so with a model with relatively few features that are easily interpreted.

- It is thus desirable to adjust R^2 to take account of the size of the model

The Adjusted R^2 Value:

ADJUST EACH TERM BY
$\div$ BY df

$df_{SSE} = n-p-1$ $df_{SST} = n-1$

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

What about Goodness-of-Fit?

- The objective of multiple linear regression is not simply to explain the most variation in the data, but to do so with a model with relatively few features that are easily interpreted.

- It is thus desirable to adjust R^2 to take account of the size of the model

$p \uparrow \quad (n-p-1) \downarrow$

$SSE/(n-p-1) \uparrow$

The Adjusted R^2 Value:

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

What about Goodness-of-Fit?

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

Which Features Should we Keep?

Model Selection:

- Try all possible combinations of p features and choose the best combo (terrible idea)

2^p possible models

$p = 30$, $2^{30} = 1{,}073{,}741{,}824$

Which Features Should we Keep?

Model Selection:

- **Forward Selection**: A greedy algorithm for adding features

    1. Fit model with an intercept but no slopes
    2. Fit p-SLR models, 1 for each feature. Add the one that improves performance the most based on some measure (e.g. SSE or F-statistic)
    3. Fit (p-1)-MLR models, 1 for each remaining feature. Add the one that improves performance the most
    4. Repeat until some stopping criterion is reached

Which Features Should we Keep?

Model Selection:

- **Backward Selection**: A greedy algorithm for removing features

    1. Fit model with all available features

    2. Remove feature with largest p-value (least-significant feature)

    3. Repeat until some stopping criterion is reached

Tutorial Problem Quiz!

1. **Advertising**: I want to know if the set of {news, radio} have slope parameters that are significantly different from zero. What test should I use?

   PARTIAL F-TEST W/ NEWS & RADIO AS SUBSET

2. **Home Prices**: I have n=1000 data points and 30 features. I want to learn the 10 best features to use in a predictive model. How should I find them?

   BACKWARD OR FORWARD SELECTION

3. **Home Prices**: I have n=100 data points and 200 features. I want to learn the 20 best features to use in a predictive model. How should I find them?

   FORWARD SELECTION

4. **Shark Attacks**: I have n=50 days of data on shark attacks and have constructed an MLR model based on 20 features. I want to measure how good my model is. What should I use?

   ADJUSTED R^2 -VALUE

OK! Let's Go to Work!

Get in groups, get out laptop, and open the Lecture 23 In-Class Notebook

Let's:

- See the Problem of Multiple Comparisons in practice!
- Use Backward Selection to determine which polynomial terms we need in a polynomial regression model.