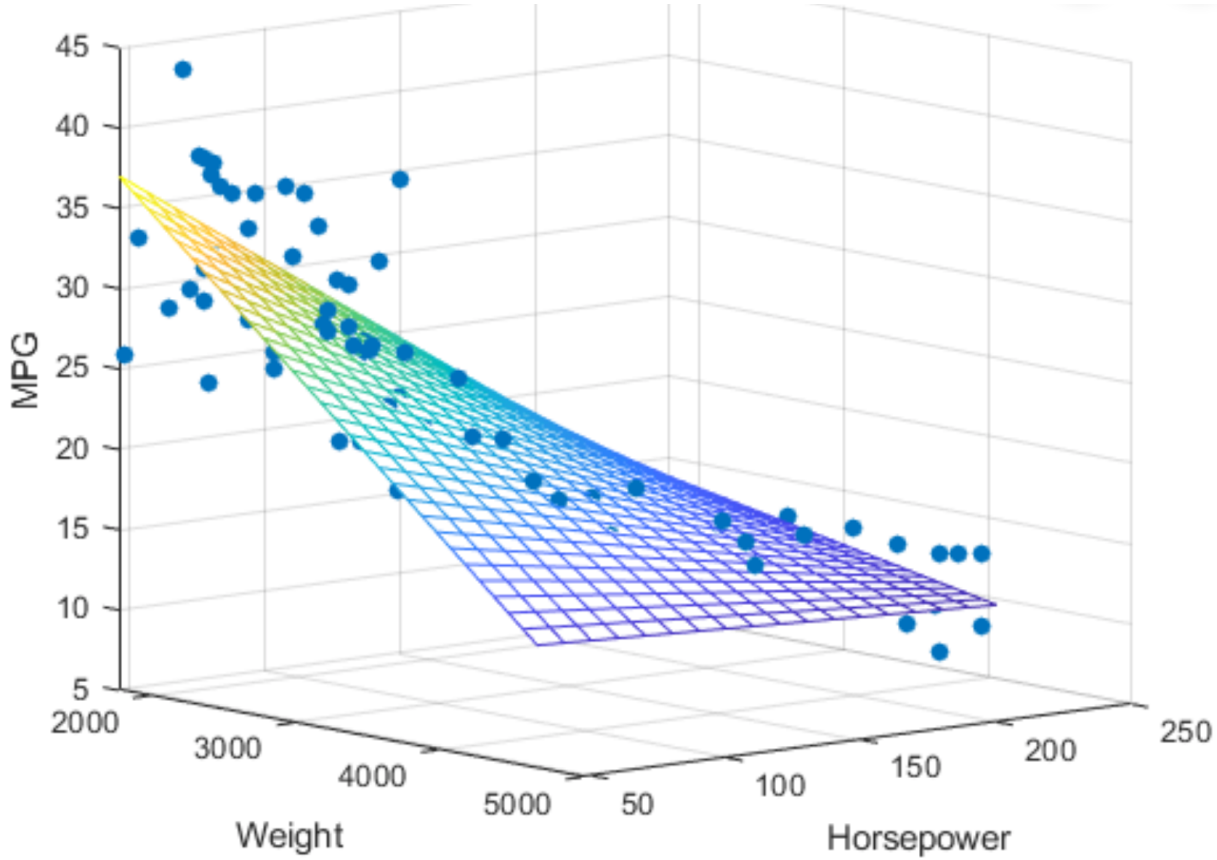
CSCI 3022: Intro to Data Science

Lecture 23: Inference and Model Selection in Multiple Linear Regression

Rachel Cox

Department of Computer Science



Announcements & Reminders

- HW5 - Due Friday April 17th at 11:59 pm

What will we learn today?

- ❑ Multiple Linear Regression
- ❑ F-statistic
- ❑ Forward selection & Backward selection

- ❑ *Introduction to Statistical Learning, Chapter 3*

Workflow for Simple Linear Regression

1) Plot the data as a scatter plot

   Does linearity seem appropriate?

   Compute $\hat{\beta}_0, \hat{\beta}_1$ and overlay the best-fit line $y = \hat{\beta}_0 + \hat{\beta}_1 X$

$y = \hat{\alpha} + \hat{\beta} x$

2) Consider assumptions of SLR:

   Plot a histogram of the residuals: are they normal?

   Plot the residuals against x: are they changing?

3) Perform inference (e.g. on βs)

$$\begin{cases} H_0: \beta = 0 \\ H_1: \beta \neq 0 \end{cases}$$

stats. linregress

Multiple Linear Regression (MLR)

MLR is like SLR, but now we can attempt to predict Y with a variety of things, including both different features/predictors X as well as transformations and augmentations of the original variables, like using x^2.

example in notebook 22

Process:

- Plot the linear model.
- See if some predictors are redundant.
- Plot residuals of the linear model.
- Analyze and use various techniques to refine the model.

Multiple Linear Regression (MLR)

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 can we reduce to just a few?
- How well does the model fit the data? How well does just a subset of features do?

Is at least one feature important?

- In the SLR setting, we can do a hypothesis test to determine if $\beta_1 = 0$

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

- In the MLR setting with p features, we need to check whether ALL coefficients are 0:

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

This would indicate no relationship between the features and Response.

H_1: $\beta_k \neq 0$ for at least one value of k in 1, 2, … , p

→ At least one feature does have a relationship with the Response

- The null hypothesis in the MLR case says that there is no useful linear relationship between y and any of the p predictors.

- This test is based on a statistic that has an F distribution when H_0 is true.

computing an F statistic – this is how we carry the hypothesis test for MLR

The F–test

The **F-test**:

We test the hypothesis via the **F-statistic**: $F = \frac{\frac{SST - SSE}{p}}{\frac{SSE}{n-p-1}}$

p = # features

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

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

- The F-statistic is a measure of how much better our model is than just using the mean.

The F-test

F statistic

The **F-test**: $F = \frac{\frac{SST - SSE}{p}}{\frac{SSE}{n-p-1}}$.

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

Suppose H_0 were true. What would F be?

$F \approx 1$

Suppose H_1 were true. What would F be?

→ Better explanation
→ Lower SSE

} SST - SSE is larger ⇒ F is higher (than 1)

Hypothesis testing:

pvalue = 1 − stats.f.cdf($F_{statistic}$, p, n−p−1)

$F \geq F_{\alpha, p, n-p-1}$ If F is ≥ threshold, reject null hypothesis.

p ↑ # features

The partial F–test

- Full model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
- Reduced model: $y = \beta_0 + \beta_2 x_2 + \beta_4 x_4$

Are the missing features important, or are we okay going with the reduced model?

We can answer this with a Partial F-test!

H_0: $\beta_1 = \beta_3 = 0$
H_1: at least one of $\beta_1, \beta_3 \neq 0$

Strategy: Fit the full and reduced models. Determine if the difference in performance is real or just due to chance.

The partial F–test

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

Test statistic: $F = \frac{(SSE_{reduced} - SSE_{full})/(p-k)}{SSE_{full}/(n-p-1)} \sim F_{p-k,n-p-1}$

$\{ df_{SSE_{Red}} - df_{SSE_{Full}}$

$= (n-1-k) - (n-p-1)$

$= -k+p$

$= p-k$

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

Why use the F-tests?

notebook 22

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 is the null hypothesis is in fact true?
  - If we had 100 parameters, about 5 would be significant just by chance.
  - **Problem of Multiple Comparisons**

```
In [6]: model.summary()
```

Out[6]:

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, 10 Jul 2018	**Prob (F-statistic):**	1.58e-96
Time:	18:04:05	**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

Quantifying model goodness–of–fit

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

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ \ldots + \hat{\beta}_p x_p \right) \right)^2$$

Like in SLR, the MLR total sum of squares, SST, is:

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

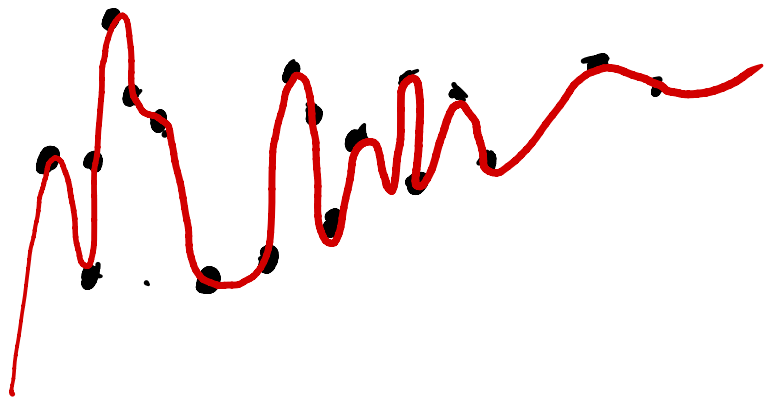
The coefficient of determination, R^2, is: $\boldsymbol{R^2 = 1 - \frac{SSE}{SST}}$

Quantifying model goodness–of–fit

Problem: The standard R^2 value can be artificially inflated by adding lots and lots of frivolous features. (You can fit anything with a polynomial of high enough degree.)

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

x_1 – the interior size of the house.
x_2 – the size of the lot
x_3 – the number of bedrooms
x_4 – the number of bathrooms
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

Quantifying model goodness–of–fit

Principle of Parsimony -- The principle that the most acceptable explanation of an occurrence, phenomenon, or event is the simplest, involving the fewest entities, assumptions, or changes.

- The objective of MLR 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 account for the size of the model (i.e. # of features)

R_a^2

Quantifying model goodness-of-fit

AIC Akaike information criterion
BIC Bayesian information criterion

$R^2 = 1 - \frac{SSE}{SST}$, but let's adjust each of SSE and SST by their degrees of freedom.

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

The **adjusted R^2** value is:

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

```
In [8]: model.summary()
```

Out[8]:

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:	Sun, 17 Nov 2019	**Prob (F-statistic):**	1.58e-96
Time:	14:57:30	**Log-Likelihood:**	-386.18
No. Observations:	200	**AIC:**	780.4
Df Residuals:	196	**BIC:**	793.6
Df Model:	3		
~~**Covariance Type:**~~	~~nonrobust~~		

data points

features

hypothesis test p-value

Model selection: which features should we keep?

Here's an idea: Try all possible combinations of p features, and choose the best combo.

- There are $\sum_{i=0}^{p} \binom{p}{p-i} = 2^p$ possible models
- With $p = 30$, that's about $2^{30} \approx 1{,}073{,}741{,}824$ models to test.

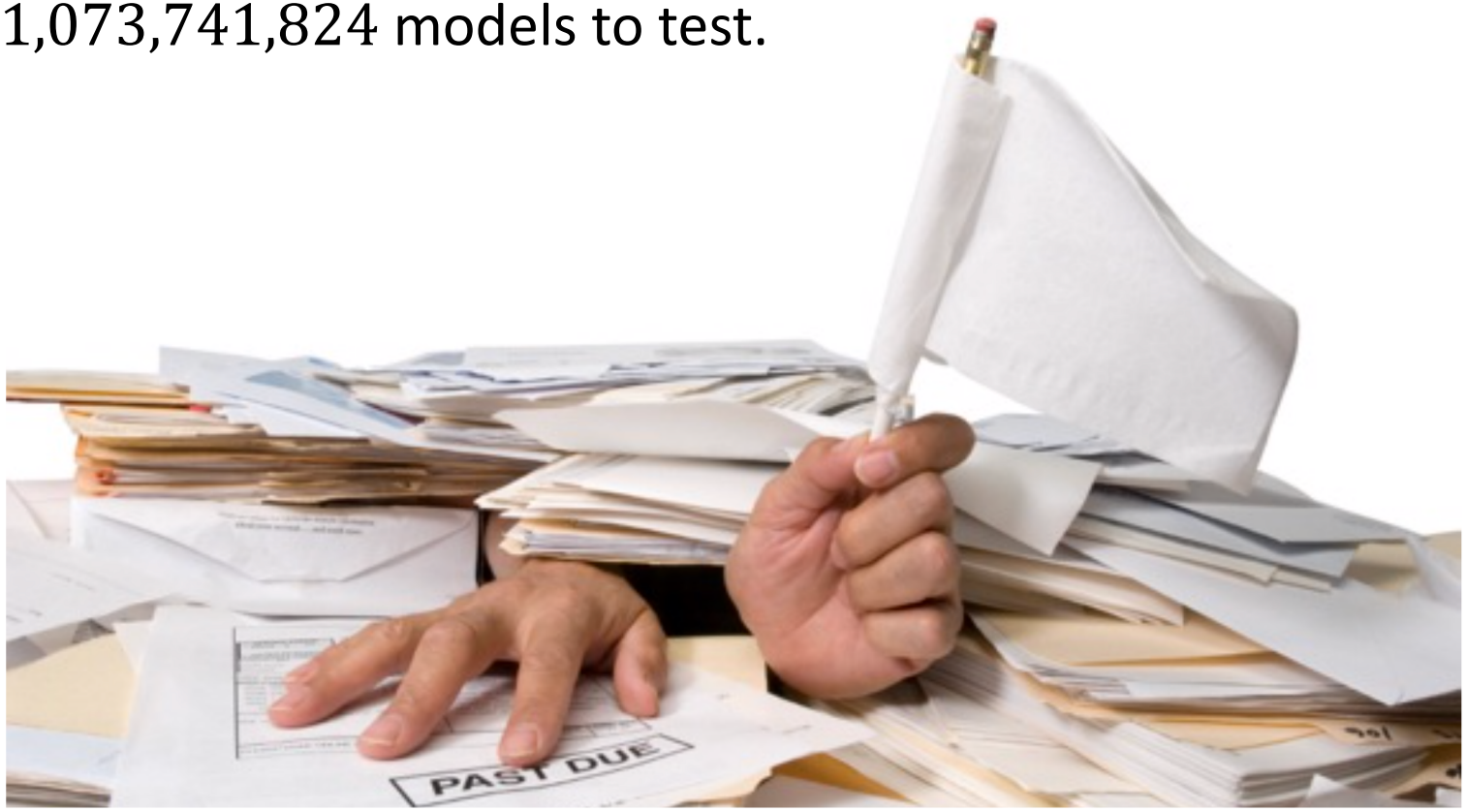
e.g. X_1, X_2, X_3, X_4

X_1, X_2, X_3 X_4

X_1, X_2 X_1, X_3

X_1, X_4 X_2, X_3

X_2, X_4 X_3, X_4



Model selection: which features should we keep?

Forward selection: A greedy algorithm for adding features.

1) Fit a null model with an intercept but no slopes $y = \beta_0$
2) Fit p individual SLR models -- 1 for each possible feature. Add to the null model the one that improves the performance the most, based on some measure. (e.g. decreases SSE the most, or increases F-statistic the most).
3) Fit $p - 1$ MLR models -- 1 for each of the remaining features along with the feature you isolated from step 2. Add the one that improves model performance the most.
4) Repeat until some stopping criterion is reached. (e.g. some threshold SSE, or some fixed number of features.)

Model selection: which features should we keep?

Backward selection: A greedy algorithm for removing features.

1) Fit model with all available features
2) Remove the feature with the largest p-value (i.e. the least significant feature)
3) Repeat until some stopping criterion is reached. (e.g. some threshold SSE, or some fixed number of features.)

Next Time:

- MLR and ANOVA

Analysis of Variance