Multiple Regression: A Model for the Mean

EXTRA SUMS OF SQUARES TESTS

TESTING DIFFERENT HYPOTHESES

A PRELIMINARY LOOK AT PREDICTION

Testing For Groups of Explanatory Variables

Extra Sums of Squares F-Test

Consider comparing the "reduced" model:

$$\mu\{salePrice|ft^2, zipcode, nBedrooms, nBaths\}$$
$$= \beta_0 + \beta_1 ft^2 + \beta_2 zipcode + \beta_3 nBedrooms + \beta_4 nBaths$$

with a "full" model that interacts each of the main effects with zipcode:

$$\mu\{salePrice|ft^2, zipcode, nBedrooms, nBaths\}$$
$$= \beta_0 + \beta_1 ft^2 + \beta_2 zipcode + \beta_3 nBedrooms + \beta_4 nBaths +$$
$$\beta_5 ft^2 * zipcode + \beta_6 nBedrooms * zipcode + \beta_7 nBaths * zipcode$$

If we want to test for the suitability of the more complex model, we need to test the hypothesis:

H_0: β_5 = β_6 = β_7 = 0

H_A: At least one coefficient $\neq 0$

The default t-tests do not address this hypothesis

Extra Sums of Squares F-Test: General Case

(see lecture notes: 5c_ESSforSpock for a discussion of extra sums of squares tests for ANOVA and Chapter 10.3 in the book)

```
PROC GLM DATA = housing PLOTS=all;
    CLASS zipCode (ref = '75224');
    MODEL salePrice = sqFootage zipCode nBedrooms nBathrooms;
RUN;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	3.988723E13	9.9718076E12	88.94	<.0001
Error	24	2.6908501E12	112118754091		
Corrected Total	28	4.257808E13			

$H_0: \beta_5 = \beta_6 = \beta_7 = 0$

H_A: At least one coefficient $\neq 0$

```
PROC GLM DATA = housing PLOTS=all;
    CLASS zipCode (ref = '75224');
    MODEL salePrice = sqFootage zipCode nBedrooms nBathrooms
                      sqFootage*zipCode nBedrooms*zipCode nBathrooms*zipCode;
RUN;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	4.0733679E13	5.8190971E12	66.26	<.0001
Error	21	1.8444011E12	87828624088		
Corrected Total	28	4.257808E13			

```
DATA pval;
   p = CDF('F',((2.6908501E12 - 1.8444011E12)/(24-21))/87828624088, 24 - 21, 21);
RUN;

PROC PRINT DATA = pval;
RUN;
```

Obs	p
1	0.95623

P-value = 0.0437

What does this mean?

Extra Sums of Squares F-Test: Special Cases

There are two special cases of note:

- Testing a single coefficient
- Testing the fit of a model to the intercept-only model

We will cover each of these in the next slides..

Testing a Single Coefficient

Compare the “reduced” model:

$$\mu\{salePrice|ft^2, zipcode, nBedrooms, nBaths\}$$
$$= \beta_0 + \beta_1 ft^2 + \beta_2 zipcode + \beta_3 nBedrooms + \beta_4 nBaths$$

to the “full” model that adds an interaction of nBaths and zipcode:

$$\mu\{salePrice|ft^2, zipcode, nBedrooms, nBaths\}$$
$$= \beta_0 + \beta_1 ft^2 + \beta_2 zipcode + \beta_3 nBedrooms + \beta_4 nBaths + \beta_5 nBaths * zipcode$$

This reduces to testing

$H_0: \beta_5 = 0$
$H_A: \beta_5 \neq 0$

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	219885.216	B	338511.8432	0.65	0.5224	-480379.884	920150.317
sqFootage	221.119		93.2692	2.37	0.0265	28.177	414.061
zipcode 75225	-1075799.128	B	381054.9625	-2.82	0.0096	-1864071.376	-287526.880
zipcode 75224	0.000	B	.	.	.	.	.
nBedrooms	-105504.679		107916.4652	-0.98	0.3384	-328746.896	117737.538
nBathrooms	-19026.733	B	150513.2001	-0.13	0.9005	-330387.010	292333.543
nBathrooms*zipcode 75225	437258.280	B	140237.3223	3.12	0.0048	147155.276	727361.284
nBathrooms*zipcode 75224	0.000	B	.	.	.	.	.

```
PROC GLM DATA = housing PLOTS=all;
    CLASS zipCode (ref = '75224');
    MODEL salePrice = sqFootage zipCode nBedrooms nBathrooms nBathrooms*zipCode/ SOLUTION CLPARM;
RUN;
```

(Exercise: Convince yourself you get the same answer as an extra sums of squares F-test)

Testing the Fit of a Model to the Intercept-only Model

The default PROC GLM output provides this test:

```
PROC GLM DATA = housing PLOTS=all;
    CLASS zipCode (ref = '75224');
    MODEL salePrice = sqFootage zipCode nBedrooms nBathrooms;
RUN;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	3.988723E13	9.9718076E12	88.94	<.0001
Error	24	2.6908501E12	112118754091		
Corrected Total	28	4.257808E13			

Exercise: Convince yourself that this output corresponds to an extra sums of squares test for the considered model vs. the "reduced" model

$$\mu\{salePrice|ft^2, zipcode, nBedrooms, nBaths\} = \beta_0$$

Making Predictions

Returning to Some Scientific Questions

- It is very important to do statistics from the perspective of a scientist
- This means always keeping in mind the scientific question(s) of interest
- What are some scientific questions of interest for the housing data?

- Suppose instead of trying to infer the relationship between aspects of a house and its sales price, we want to get a prediction of the sales price.

A survey of housing data + multiple regression could be used

Prediction

- Prediction is a very important task that seems very similar to inference
  - Here, we want to estimate a function $\hat{\mu}\{Y|X\}$ such that I can produce an ”estimate” or “prediction” of Y with only observing X
- However, the details and motivation are quite different
- Primarily:
  - Confounding is not a concern when building a predictive model
  - Confounding is a primary concern when attempting to do inference
- What is a good example that demonstrates the difference between building a predictive model and an inferential model?

Prediction

- Suppose that the true relationship between Y and X is given by a function $\mu\{Y|X\}$ and some random error ε:

$$Y = \mu\{Y|X\} + \varepsilon$$

- Then the quality of our prediction can be decomposed into three quantities:
  - Approximation error: $\mu\{Y|X\} - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right)$
  - Estimation error: $(\beta_0 + \sum_{j=1}^{p} \beta_j x_j) - (\widehat{\beta_0} + \sum_{j=1}^{p} \widehat{\beta}_j x_j)$
  - Irreducible error: $\mu\{Y|X\} - \varepsilon$
- The best possible prediction is given by $\mu\{Y|X\}$, which is unknown
- A good prediction $\hat{\mu}\{Y|X\}$ has small Approximation and Estimation errors

Returning to: Including "Wrong" Explanatory Variables

There are roughly four possible (unknown) outcomes:

1. Model is correctly specified. All relevant explanatory variables are included
2. Model is under specified. Some important explanatory variables are omitted
3. Model includes unimportant explanatory variables
4. Model is over specified. Some redundant (or nearly redundant) explanatory variables are included

1. Ideal case. Estimates are unbiased
2. Estimates are biased and we are overestimating σ. We can get incorrect inferences due to confounding
3. Estimates are unbiased. However, the inclusion of irrelevant parameters decreases power
4. We'll return to this during multicollinearity

Making Predictions

When the scientific question of interest is primarily about making predictions, the decrease in power from too large a model is important

If we want to build a predictive model for housing prices, confounding isn't a concern

Let's return to the "main effects" fitted model

Making Predictions

$$\hat{\mu}\{salePrice|ft^2, zipcode, nBedrooms, nBaths\}$$
$$= \widehat{\beta_0} + \widehat{\beta_1}ft^2 + \widehat{\beta_2}zipcode + \widehat{\beta_3}nBedrooms + \widehat{\beta_4}nBaths$$

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	-413793.0938	B	316100.7504	-1.31	0.2029	-1066192.978	238606.7902
sqFootage	252.8059		108.2575	2.34	0.0282	29.3733	476.2384
zipcode 75225	-77819.1511	B	241444.0896	-0.32	0.7500	-576135.2603	420496.9581
zipcode 75224	0.0000	B	.	.	.	.	.
nBedrooms	-181430.9105		122758.9588	-1.48	0.1524	-434792.9490	71931.1280
nBathrooms	380639.9690		92117.7525	4.13	0.0004	190518.2722	570761.6658

- Previously, we left the zipcode and nBedroom terms in the model, due to being more concerned about confounding
- When considering building a predictive model, it makes sense to make sure that each of the terms in the model are important
- In this case, we can try and jointly test for inclusion of zipcode and nBedroom

Making Predictions

$$\hat{\mu}\{salePrice|ft^2, zipcode, nBedrooms, nBaths\} = \widehat{\beta_0} + \widehat{\beta_1}ft^2 + \widehat{\beta_2}zipcode + \widehat{\beta_3}nBedrooms + \widehat{\beta_4}nBaths$$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	3.988723E13	9.9718076E12	88.94	<.0001
Error	24	2.6908501E12	112118754091		
Corrected Total	28	4.257808E13			

Fail to reject…
What does that mean?

Source	DF	SS	MS	F	Pr > F
Model	2	2.7E11	1.08E11	0.97	0.394
Error	24	2.69E12	1.12E11		
Corrected Total	26	2.96E12			

$$\hat{\mu}\{salePrice|ft^2, zipcode, nBedrooms, nBaths\} = \widehat{\beta_0} + \widehat{\beta_1}ft^2 + \widehat{\beta_2}nBaths$$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3.9622781E13	1.9811391E13	174.30	<.0001
Error	26	2.955299E12	113665344887		
Corrected Total	28	4.257808E13			

Making Predictions

```
DATA prediction;
    INPUT salePrice sqFootage nBedrooms nBathrooms zipcode $;
    DATALINES;
    . 2500 3 2 75224
;

DATA prediction;
    SET prediction housing;

PROC GLM DATA = prediction;
    CLASS zipCode (ref = '75224');
    MODEL salePrice = sqFootage nBathrooms / CLI CLM;
RUN;
```

Observation		Observed	Predicted	Residual	95% Confidence Limits for Mean Predicted Value	
1	*	.	336342.847	.	117097.203	555588.492
2		2250000.000	2558628.486	-308628.486	2291147.232	2826109.740
3		2650000.000	2611165.815	38834.185	2299662.172	2922669.458

$$\hat{\mu}\{salePrice|ft^2, nBaths\} = \widehat{\beta_0} + \widehat{\beta_1}ft^2 + \widehat{\beta_2}nBaths = \$336342.84$$

The confidence limits we choose again depends on the scientific question..

(details in 7b_linearRegressionDarren)

(This discussion is from Chapter 10.2.4. We will return to this again in Chapter 12 when we discuss model selection)