

Confounding and Interaction

Applied Regression and Other Multivariable Methods
Sections 11-1 - 11-7

Overview

- There are two primary goals of regression
 - Predicting the dependent variable from set of predictors
 - Quantifying the relationship between predictors and dependent variable
- So far, have primarily dealt with prediction
 - High overall R^2 balanced against model simplicity
 - Low MSE balanced against residual degrees of freedom
 - Strength of relationship after adjusting for other variables
- Recall Topic 9
 - Discussed meaning of parameter estimates after adjusting for others (SMK after adjusting for AGE)
 - Discussed interaction in terms of non-parallel regression lines (SBP vs QUET at each level of SMK)
- Will now discuss these ideas in more detail

Confounding and Interaction

- Want to account for other nuisance variables
- These variables known as extraneous/control variables or covariates
- Goal: Interest in assessing an association
- Consider variables Y , X , and covariate Z
- Want to answer
 1. Is the estimate of association between Y and X dramatically different if one adjusts or does not adjust for Z ?
 2. Is the estimate of association between Y and X meaningfully different at different values of Z ?
- #1 deals with confounding
- #2 deals with interaction

Confounding

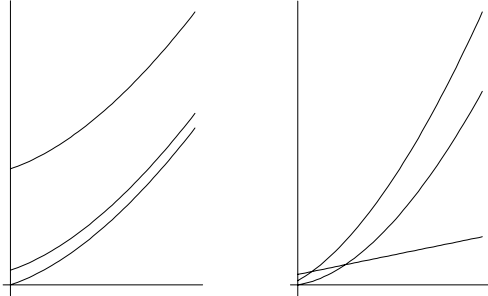
- Estimate of association usually the parameter est
- Could also consider the (partial) correlation coef
- Confounding causes spurious results
- To assess confounding, compare the coefficient on X for the following two models
$$Y = \beta_0 + \beta_1 X \quad \text{and} \quad Y = \beta_0^* + \beta_1^* X + \beta_2^* Z$$
- If $\hat{\beta}_1$ is "different" than $\hat{\beta}_1^*$, then confounding is present and Z should be included in the final model to avoid spurious results
- In question 3 (Chpt 10) spurious results are discussed in terms the partial correlation coefficient. The strength of the relationship dramatically changes after adjusting for confounders.
- This approach does not require a statistical test. Simply a comparison of adjusted vs unadjusted estimates.
- Not the same as testing $H_0 : \beta_1^* = 0$ or $\rho_{YX|Z} = 0$

Interaction

- An interaction is when the relationship between Y and X is different at different values of Z .
- For any two values of X (x_1, x_2) and Z (z_1, z_2) is

$$\mu_{Y|X=x_2, Z=z_1} - \mu_{Y|X=x_1, Z=z_1} \neq \mu_{Y|X=x_2, Z=z_2} - \mu_{Y|X=x_1, Z=z_2}?$$

- Is regression of Y and X for different Z parallel?
- Recall relationship is not necessarily linear with X



13-4

Interaction

- To assess interaction first describe the relationship
- Mathematically describe it using product (XZ)

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ$$

– If $Z = 1$

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \beta_2 + \beta_3 X \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X \end{aligned}$$

– If $Z = 2$

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + 2\beta_2 + 2\beta_3 X \\ &= (\beta_0 + 2\beta_2) + (\beta_1 + 2\beta_3) X \end{aligned}$$

- Both the slope and intercept of Y vs X have changed
- Change in slope due to XZ term
- Then do a statistical test to see if XZ is important
- Adjusting for a confounder can sometimes hide interaction. Thus always adjust for interaction first.
- If interaction present, cannot simply compare un-adjusted and adjusted estimates

13-5

Example Problem

Consider the hypothetical sample of 32 white males over the age of 40 from the town of Angina that was described in Chapter 5 (Prob #2). Suppose we're interested in describing the relationship between systolic blood pressure (sbp) and body type (quet) and have two possible covariates, past history of smoking (smk) and age (age).

The first step in the process is to assess whether there is any interaction between the covariates and independent variable of interest.

I will look at this three ways

1. Looking at just the SMK covariate and interaction
2. Looking at just the AGE covariate and interaction
3. Looking at both AGE and SMK and their interactions

```
options nocenter; options reset=global colors=(none);
data problem81;
infile 'i:\www\datasets502\EX0502.DAT' firstobs=2 dlm='09'x;
input person sbp quet age smk;
quet_smk = quet*smk; quet_age = quet*age;

proc reg corr;
model sbp = quet smk quet_smk;
model sbp = quet age quet_age;
model sbp = quet smk age quet_smk quet_age quet_smk / pcorr1;
run;
```

13-6

Correlation

Variable	quet	smk	quet_smk	sbp	age	quet_age
quet	1.0000	-0.0714	0.1191	0.7420	0.8028	0.9512
smk	-0.0714	1.0000	0.9732	0.2473	-0.1395	-0.1053
quet_smk	0.1191	0.9732	1.0000	0.3717	-0.0128	0.0624
sbp	0.7420	0.2473	0.3717	1.0000	0.7752	0.8095
age	0.8028	-0.1395	-0.0128	0.7752	1.0000	0.9429
quet_age	0.9512	-0.1053	0.0624	0.8095	0.9429	1.0000

There does not appear to be any dangerous correlations between predictors except between (quet_smk and smk), (quet_age and age), and (quet_age and quet). We will ignore this for now. In Chapter 12, we will discuss a way to avoid confounding interaction terms with the main factor terms. This method is known as centering and involves subtracting off the mean.

13-7

Looking at SMK and QUET_SMK

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4184.10759	1394.70253	17.42	<.0001
Error	28	2241.86116	80.06647		
Corrected Total	31	6425.96875			

Root MSE	8.94799	R-Square	0.6511
Dependent Mean	144.53125	Adj R-Sq	0.6137
Coeff Var	6.19104		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	49.31176	19.97235	2.47	0.0199
quet	1	26.30283	5.70349	4.61	<.0001
smk	1	29.94357	24.16355	1.24	0.2256
quet_smk	1	-6.18478	6.93171	-0.89	0.3799

- When SMK=0

$$SBP = 49.31 + 26.30QUET + 0 + 0$$

$$= 49.31 + 26.30QUET$$

- When SMK=1

$$SBP = 49.31 + 26.30QUET + 29.94 - 6.18QUET$$

$$= 79.25 + 20.13QUET$$

This relationship is demonstrated in slide 9-17. Notice that the SMK=1 line is flatter than the SMK=0 line so that the distance between the two gets narrower as QUET increases. This suggests the possibility of an interaction but is there statistical significance? The P-value of the interaction term is only 0.38 so there is not enough evidence to suggest an interaction.

13-8

Looking at AGE and QUET_AGE

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4354.47366	1451.49122	19.62	<.0001
Error	28	2071.49509	73.98197		
Corrected Total	31	6425.96875			

Root MSE	8.60128	R-Square	0.6776
Dependent Mean	144.53125	Adj R-Sq	0.6431
Coeff Var	5.95115		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	207.36956	86.36535	2.40	0.0232
quet	1	-34.11701	25.21679	-1.35	0.1869
age	1	-1.84682	1.66861	-1.11	0.2778
quet_age	1	0.82239	0.46253	1.78	0.0863

Since AGE is continuous, it is not as easy to assess the interaction. However, if we look at two particular values of AGE

- When AGE=50

$$SBP = 207.37 - 34.12QUET - 1.85(50) + 0.82(50)QUET$$

$$= 114.87 + 6.88QUET$$

- When AGE=60

$$SBP = 207.37 - 34.12QUET - 1.85(60) + 0.82(60)QUET$$

$$= 96.37 + 15.08QUET$$

This suggests the possibility of an interaction but is there statistical significance? The P-value of the interaction term is 0.086 which is marginally significant. It is saying that as AGE increases, a change in QUET has a larger effect. This seems reasonable.

13-9

Looking at both AGE and SMK

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	5081.86742	1016.37348	19.66	<.0001
Error	26	1344.10133	51.69621		
Corrected Total	31	6425.96875			

Root MSE	7.19001	R-Square	0.7908
Dependent Mean	144.53125	Adj R-Sq	0.7506
Coeff Var	4.97471		

Parameter Estimates						Squared Partial
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Corr Type I
Intercept	1	185.69424	73.70952	2.52	0.0182	.
quet	1	-32.41001	21.84299	-1.48	0.1499	0.55057
smk	1	3.46636	20.61748	0.17	0.8678	0.20167
age	1	-1.35438	1.40388	-0.96	0.3436	0.33373
quet_age	1	0.73888	0.38735	1.91	0.0676	0.12193
quet_smk	1	1.80295	5.95779	0.30	0.7646	0.00351

Should we include quet_smk?

$$F = \frac{.00351/1}{(1 - .00351)/26} = 0.092$$

This gives a P-value of 0.7646 because it is the same test as the parameter test given in the output.

Should we include quet_age after throwing out quet_smk?

$$F = \frac{.12193/1}{(1 - .12193)/27} = 3.749$$

This gives a P-value around 0.063 which is very close to being significant. I'd likely leave it in the model.

13-10

Example Problem

The next step is to assess confounding. If quet_age is included in the model, then we cannot simply compare the parameter estimates of QUET to see if AGE is a confounder. To simplify the example here, we will assume there is no interaction present (i.e., no quet_age in the model because it is non-significant).

In order to make this assessment, we will look at the following models.

```
proc reg;
  model sbp = quet;
  model sbp = quet age;
  model sbp = quet smk;
  model sbp = quet age smk;
run;
```

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3537.94574	3537.94574	36.75	<.0001
Error	30	2888.02301	96.26743		
Corrected Total	31	6425.96875			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	70.57640	12.32187	5.73	<.0001
quet	1	21.49167	3.54515	6.06	<.0001

Here $\beta_1 \approx 21.5$

13-11

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4120.59225	2060.29612	25.92	<.0001
Error	29	2305.37650	79.49574		
Corrected Total	31	6425.96875			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	55.32344	12.53475	4.41	0.0001
quet	1	9.75073	5.40246	1.80	0.0815
age	1	1.04516	0.38606	2.71	0.0113

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4120.36649	2060.18325	25.91	<.0001
Error	29	2305.60226	79.50353		
Corrected Total	31	6425.96875			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	63.87603	11.46811	5.57	<.0001
quet	1	22.11560	3.22996	6.85	<.0001
smk	1	8.57101	3.16670	2.71	0.0113

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4889.82570	1629.94190	29.71	<.0001
Error	28	1536.14305	54.86225		
Corrected Total	31	6425.96875			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	45.10319	10.76488	4.19	0.0003
quet	1	8.59245	4.49868	1.91	0.0664
age	1	1.21271	0.32382	3.75	0.0008
smk	1	9.94557	2.65606	3.74	0.0008

13-12

13-13

- When AGE is added to the model, the coefficient on QUET drops down to 9.75. This is quite a large jump suggesting that there is some confounding and we need to control for AGE.
- When SMK is added to the model, the coefficient on QUET increases to 22.11. This is not that big a jump.
- When both AGE and SMK are fit in the model, the coefficient on QUET is 8.6. This suggests that the QUET relationship should be controlled for both AGE and SMK.
- Both AGE and SMK are significant (P-value < 0.05)
- Also notice that in terms of precision of the estimate β_1 , adding SMK to the model that already has AGE reduces the std error to approximately 4.5.

If quet_age is in the model, then only SMK can be investigated as a confounder. If this model is fit, you could investigate changes in the coefficient for a specific age value.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	5077.13311	1269.28328	25.41	<.0001
Error	27	1348.83564	49.95688		
Corrected Total	31	6425.96875			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	181.70377	71.29001	2.55	0.0168
quet	1	-30.69963	20.74113	-1.48	0.1504
age	1	-1.38470	1.37654	-1.01	0.3234
quet_age	1	0.73724	0.38074	1.94	0.0634
smk	1	9.65649	2.53893	3.80	0.0007

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4354.47366	1451.49122	19.62	<.0001
Error	28	2071.49509	73.98197		
Corrected Total	31	6425.96875			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	207.36956	86.36535	2.40	0.0232
quet	1	-34.11701	25.21679	-1.35	0.1869
age	1	-1.84682	1.66861	-1.11	0.2778
quet_age	1	0.82239	0.46253	1.78	0.0863

13-14

13-15

- When AGE=50

$$SBP = 207.37 - 34.12QUET - 1.85(50) + 0.82(50)QUET$$

$$= 114.87 + 6.88QUET$$
- When AGE=50 (SMK = 0)

$$SBP = 181.70 - 30.70QUET - 1.38(50) + 0.74(50)QUET$$

$$= 112.7 + 6.30QUET$$

The change in the "coefficient" of QUET is only .5 units. Including SMK has improved the precision.

Here you can also see the difficulty in explaining confounding when interaction is present. In confounding we're trying to assess whether the coefficient on QUET changes when AGE is added to the model. However, when the interaction is included, we're saying that the SBP/QUET relationship is different at different ages. Here we see that at AGE=50, the coefficient on QUET is approximately 6.5. If we move to another AGE, the change in the coefficient will be something else. If there is an interaction, putting in an interaction term will change the relationship between the variables of interest. Thus, assessing confounding after interaction is usually irrelevant.