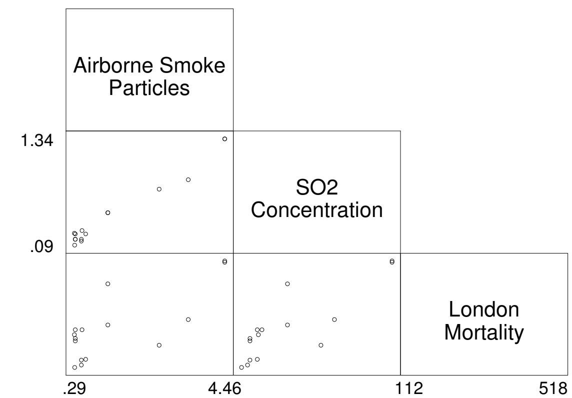
Lecture 11:
AV plots, hypothesis testing and nested models

Sandy Eckel
seckel@jhsph.edu

9 May 2008

Another Example: Mortality

Smoke, pollution & London mortality data



Mortality Example: Model

Let:

- Y = the daily mortality for London *(deaths)*
- X_1 = airborne smoke particles (mg/m^3) *(smoke)*
- X_2 = SO_2 (ppm) *(so2)*

Model:

- Systematic: $Y_i = \beta_0 + \beta_1(X_1\text{-}2) + \beta_2(X_2\text{-}.5) + \varepsilon_i$
- Random: $\varepsilon_i \sim N(0, \sigma^2)$
  - Mortality is a linear function of the concentration of airborne smoke particles *AND* the SO2 level

Mortality Example: Results

Model:
$E(\, Y \mid X \,) = \beta_0 + \beta_1(X_1\text{-}2) + \beta_2(X_2\text{-}.5)$

```
      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  2,    12) =   36.57
       Model |  205097.531     2  102548.765           Prob > F      =  0.0000
    Residual |  33654.2025    12  2804.51687           R-squared     =  0.8590
-------------+------------------------------           Adj R-squared =  0.8355
       Total |  238751.733    14  17053.6952           Root MSE      =  52.958

------------------------------------------------------------------------------
      deaths |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 smokecenter |  -220.3244   58.14314    -3.79   0.003    -347.0074   -93.64135
   so2center |   1051.816   212.5959     4.95   0.000     588.6096    1515.023
 (Intercept) |   174.7703   29.16174     5.99   0.000     111.2323    238.3083
------------------------------------------------------------------------------
```

Mortality Example
Inference: Overall F test

- Overall F-Test:
  - Are *ANY* of the covariates significant?
- H_0: $\beta_1 = \beta_2 = 0$;
- Fobs: $(2,12) = 36.57$;
- p-value = 0.0000
- Decision: At least one of the β's are nonzero

Mortality Example
Coefficient inference: individual 95% C.I. & t-tests

β_0
- $b_0 = 174.8$
  95% CI: (111.2, 238.3)
- H_0: $\beta_0 = 0$
- t_{obs}: $(12) = 5.99$
- p-value = 0.000

β_1
- $b_1 = -220.3$
  95% CI: (-347.0, -93.6)
- H_0: $\beta_1 = 0$
- t_{obs}: $(12) = -3.79$
- p-value = 0.003

β_2
- $b_2 = 1051.8$
  95% CI: (588.6, 1515.0)
- H_0: $\beta_2 = 0$
- t_{obs}: $(12) = 4.95$
- p-value = 0.000
  means p-value < 0.001

Mortality Example
Parameter Estimates Interpretation

- b_0: when smoke particles and SO_2 are at their average levels, (2 mg/m^3,and 0.5 ppm respectively), the estimated mean number of deaths is 174.8 / day
- b_1: the estimated mean mortality is 22 deaths/day lower on days when particles are 0.1 mg/m^3 higher *if SO_2 is unchanged*
- b_2 : *(You do!)*

Mortality Example
Association between x and y

- The estimate for airborne smoke particles is $b_1 = -220$, implying that smoke particles and mortality have a *negative* relationship
  - i.e. an *increase* in smoke particles is associated with a *decrease* in mortality, after adjusting for SO_2 levels.

Mortality Example
Negative Association??

- BUT WAIT!
- Look at the plot of *deaths vs smoke* presented previously. Shouldn't the relationship be *positive* instead?!
- Let's run Simple Linear Regressions (SLRs) of mortality on smoke & SO_2 and see what we get

SLR Models

- Y = the daily mortality for London *(deaths)*
- X_1 = airborne smoke particles (mg/m3) *(smoke)*
- X_2 = SO_2 (ppm) *(so2)*

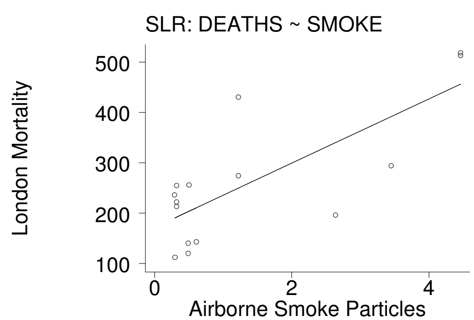
- Smoke:
  - 1) $Y_i = \beta_0 + \beta_1(X_1\text{-}2) + \varepsilon_i$
  - 2) $\varepsilon_i \sim N(0, \sigma^2)$
- SO_2:
  - 1) $Y_i = \beta_0^* + \beta_1^*(X_2\text{-}.5) + \varepsilon_i^*$
  - 2) $\varepsilon_i^* \sim N(0, \sigma^{2*})$

SLR: Deaths ~ Smoke

Parameter Estimates: $b_0 = 299.3$
$b_1 = 63.8$ *(is positive?!!)*

Amount of variation described: $R^2 = SSM / SST = 57\%$

Residual Variability left over, (undescribed by this SLR):
SSE = 1023002.216

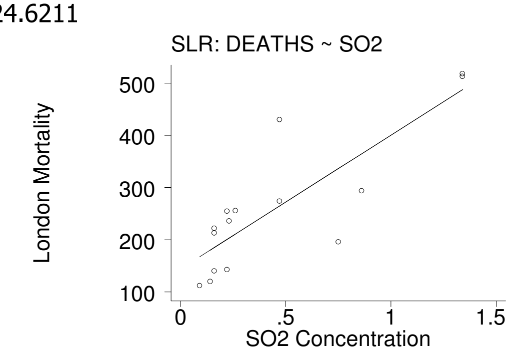


SLR: Death ~ SO_2

Parameter Estimates: $b_0 = 256.2$
$b_1 = 272.2$

Amount of variation described: $R^2 = SSM / SST = 69\%$

Residual Variability left over, (undescribed by this SLR):
SSE = 73924.6211

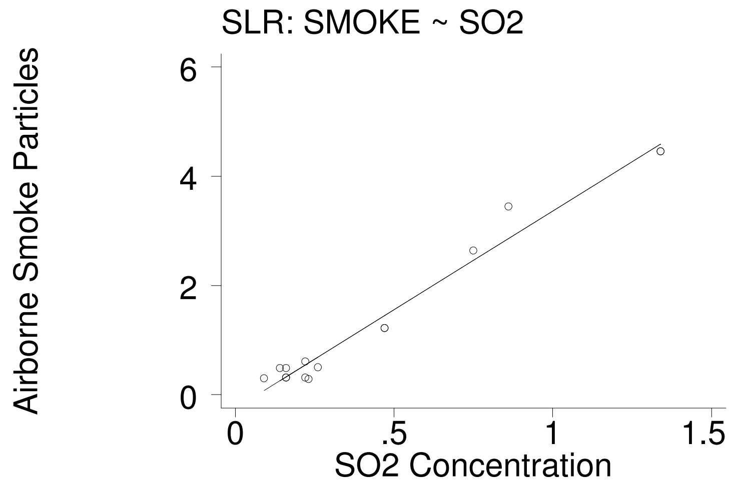


Confounding in this Example

Recall our parameter interpretations:

- β_1 = Expected change in mortality on days when particles are 0.1 mg/m^3 higher *if SO_2 is unchanged*
- Suppose we examine the relationship between smoke particle concentrations and SO_2 levels, (SLR):

SLR: Smoke ~ SO_2

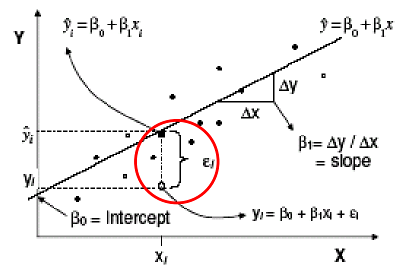


Confounding

- Smoke particle concentrations and SO_2 levels are highly related! How can we talk about *changing* smoke particle concentrations while *leaving SO_2 levels unchanged??*
- This is 'confounding'!
  - both covariates are related to the outcome and to each other
- Confounding is the reason we found differences between the SLR models and the MLR model
- We'll visualize this relationship using `Added Variable Plots'

Recall
Residuals: part "left over"

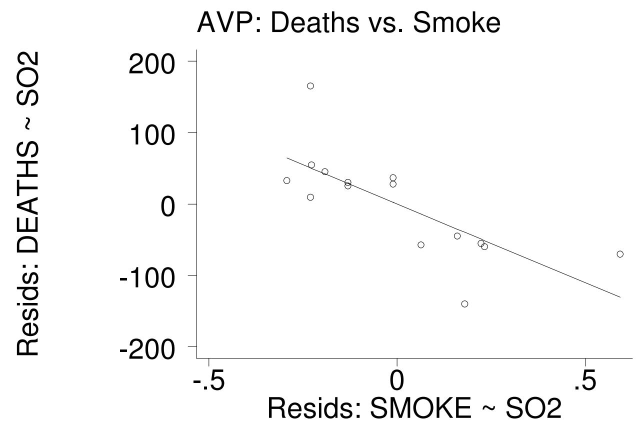
- Residuals are deviations (what's 'left over') in the response (Y) after removing what was expected given the predictor (X)
- The residuals are the part of Y that can't be predicted by X!



Adjusted Variable Plots
Idea

- Explain all the signal we can in London daily mortality using SO_2 levels
- Explain all the signal we can in smoke particle concentrations using SO_2 levels
- Explain everything that's 'left over' in mortality with everything that's 'left over' in smoke particle concentrations. The slope of this line will be the MLR coefficient!

Mortality Example
Adjusted Variable Plot



Recipe for AV plot

Recipe for obtaining the MLR slope for X_1 from an AV plot (adjusted for X_2):

1. Regress Y on X_2, save residuals as: $R_{Y|X2}$
2. Regress X_1 on X_2, save residuals as: $R_{X1|X2}$
3. Plot $R_{Y|X2}$ vs $R_{X1|X2}$ (Adjusted Variable Plot)

Regress $R_{Y|X2}$ on $R_{X1|X2}$:

$$R_{Y|X2} = \beta_0^* + \beta_1^* R_{X1|X2} + \varepsilon$$

Notes on AV Plots

- β_1^* is identical to the coefficient of X_1 from an MLR of Y on X_1 and X_2
- β_0^* (intercept) is always 0
- The AV Plot display may be misleading if Y and/or X_1 are not linearly related to the other predictors

AV Plot Recipe for Mortality Example

- Regress deaths on (centered) SO_2, save residuals
  - Removes the effects of SO2 on mortality

$$\text{Deaths} = 272 + 256\ SO_{2c} + R_{Y|X2}$$

- Regress Smoke on SO_2 (both centered), save residuals
  - Removes the effects of SO_2 on smoke particles

$$\text{Smoke}_c = -.44 + 3.6\ SO_{2c} + R_{X1|X2}$$

- Regress $R_{Y|X2}$ on $R_{X1|X2}$
  - regress deaths *adjusted for* SO_2 on smoke particles *adjusted for* SO_2
- $R_{Y|X2} = 0.0 – 220\ R_{X1|X2}$

AV plot interpretation

- Parameter from this last regression: $\beta_1^* = -220$ is the same as the related parameter from the MLR of deaths on smoke particles *and* SO_2

$$E(\text{Deaths}) = \beta_0 + \beta_1(\text{smoke-2}) + \beta_2(SO_2\text{-.5})$$
$$= 174.8 - 220\ (\text{smoke} - 2) + 1052\ (SO_2 - 0.5)$$

- This helps in our interpretations of β_1: the effect of airborne smoke particles on daily mortality after having removed (adjusted for) the effects of SO_2
  - This is what is usually meant by the term 'adjustment'

MLR and Scientific Inference

- The **single most important idea** today may be the realization that MLR can shift interpretations markedly!
- From SLR of the air pollution data:

$$E(\text{Deaths}) = 299 + 64(\text{smoke-2})$$

  - Expected deaths **increase** by an estimated 64 per mg/m^3 increase in British smoke

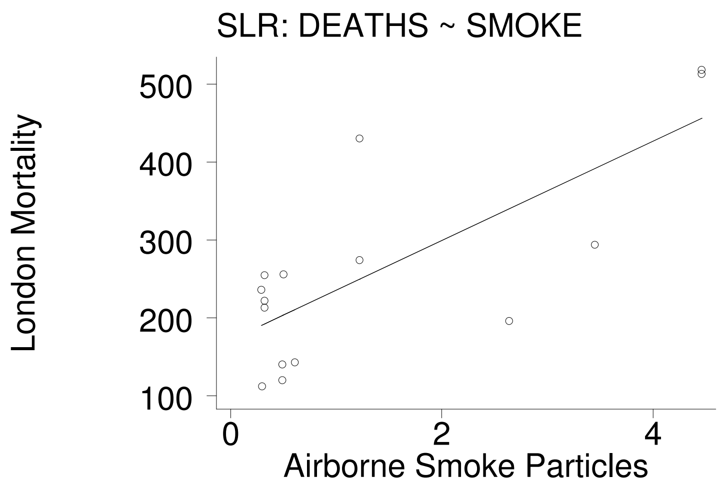
MLR and Scientific Inference

- From MLR of the air pollution data:

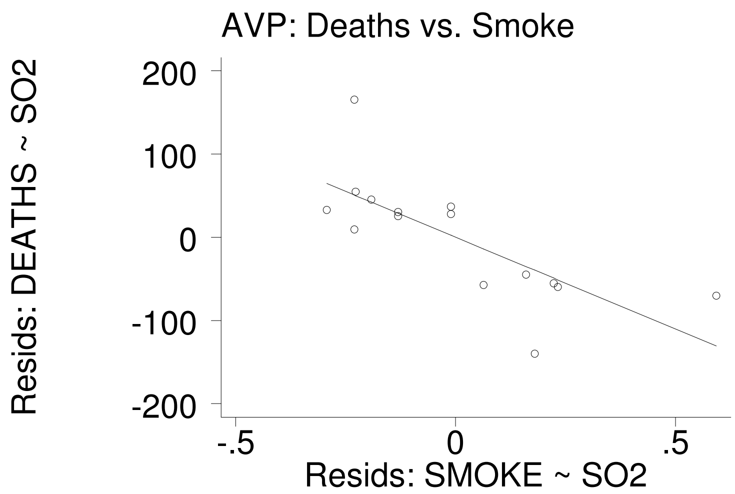
  $E(\text{Deaths}) = 174.8 - 220(\text{smoke-2}) + 1052(SO_2\text{-.5})$

  - *Controlling for SO_2*, expected deaths **decrease** 220 per mg/m^3 of British smoke
- Interpretation and value of a regression coefficient depends critically on what other variables are in the model !!

Simple Linear Regression



Multiple Linear Regression



Types of predictors in regression

- primary predictor
  - always in model
- other predictor(s)
  - can we improve prediction after adjusting for primary predictor?
  - interaction may be a component here
- potential confounder(s) (i.e., demographics)
  - only important if they change the effect of the primary predictor
  - commonly: age, gender, SES, race, etc…

Nested models

Definition: One model is nested within another if the **parent model** contains the 'original' set of variables and is *nested* within the **extended model** that contains the original set of variables plus additional variables

Nested models
Deciding whether to include variables

If the 'new variable(s)' are:

- another predictor(s)
  - assess with t-test in extended model if single variable
  - assess with F-test if two or more variables
- potential confounder(s)
  - compare CI of primary predictor in parent model to see whether new estimate of primary predictor coefficient is significantly different

Dataset

- Class health dataset
  - Outcome: number of credits
  - Primary predictor
    - housing (on or off campus)
  - Other predictors
    - health status (good/excellent or fair/poor)
    - year in school

Models

- **Parent Model (Model 1)**

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1(\text{Housing}_i)$$

1 if on-campus

0 if off-campus

credits	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
housing	.1666667	.6761572	0.25	0.807	-1.228853	1.562187
(Intercept)	16.2	.5135783	31.54	0.000	15.14003	17.25997

- **Extended Model (Model 2)**

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1(\text{Housing}_i) + \hat{\beta}_2(\text{Healthgood}_i)$$

1 if excellent/good

0 if fair/poor

credits	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
housing	.1541237	.6860262	0.22	0.824	-1.26503	1.573277
healthgood	.4139175	.7124214	0.58	0.567	-1.059838	1.887673
(Intercept)	15.9366	.6904955	23.08	0.000	14.5082	17.365

Comparing models 1 and 2

- Model 1 is nested in model 2
- Model 2 contains only one extra variable (healthgood), so use a t-test to decide whether to include healthgood
  - $p=0.567 > \alpha=0.05$ tests H_0: $\beta_2=0$
  - Fail to reject H_0
  - Conclude model 2 is no better than model 1

What if we add more than one variable?

- The t-test on each row only tests that variable *in the presence of everything else in the model*
- When more than one variable is added at a time, the t-test is not sufficient
  - The t-test only tests one variable at a time
  - Use the F-test instead to compare nested models that differ by more than one variable

When would more than one variable need to be added??

- Many modeling scenarios require adding more than one variable at once to go from the parent model to the extended model
- Commonly occurs when categorical variable needs to be added

Why do we need to specially code a categorical predictor?

- A categorical predictor (such as year in program) cannot be added as a single variable
  - If we add year (1, 2, 3, or 4) to the model in its original form, then software thinks it is a continuous predictor
  - As a continuous predictor, the difference in mean number of credits taken would be assumed to change by a constant amount for each additional year

Correct coding of a categorical predictor

- A categorical predictor should always be recoded as a set of dummy variables
  - Choose one category as the reference group
  - For each ***other*** category, create a dummy variable for membership in that category
  - You can have **`R`** do this automatically for you with the command
    `factor(mycatvar)`
    within your linear regression command

Example

- Year1 = reference group
  (no dummy variable for this group)
- **Year2** = 1 for those in year 2, 0 else
- **Year34** = 1 for those in yr 3/4, 0 else
  - very few observations, so categories were combined
- In in year 3: Year2=0, Year34=1
- For a first year: Year2=0, Year34=0

Model 3

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1(\text{Housing}_i) + \hat{\beta}_2(\text{Year2}_i) + \hat{\beta}_3(\text{Year34}_i)$$

credits	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
housing	-1.402299	.8537457	-1.64	0.115	-3.172859	.3682613
year2	.7068966	.7215468	0.98	0.338	-.7894999	2.203293
year34	-2.10197	1.087462	-1.93	0.066	-4.357228	.1532874
(Intercept)	17.34483	.9268436	18.71	0.000	15.42267	19.26698

- We cannot evaluate Year using the t-test for each row, because two variables are needed to define Year and the t tests are separate
- We must use an F-test to evaluate Year by comparing the residual sums of squares (RSS) in the parent model and in the nested model.

Comparing model RSS and Residual df

PARENT: MODEL 1

RSS_{parent} Residual df_{parent}

Source	SS	df	MS
Model	.176282088	1	.176282088
Residual	69.6333335	24	2.9013889
Total	69.8096156	25	2.79238462

Number of obs	=	26
F(1, 24)	=	0.06
Prob > F	=	0.8074
R-squared	=	0.0025
Adj R-squared	=	-0.0390
Root MSE	=	1.7033

EXTENDED: MODEL 3

$RSS_{extended}$ Residual $df_{extended}$

Source	SS	df	MS
Model	19.9853465	3	6.66178216
Residual	49.8242691	22	2.26473951
Total	69.8096156	25	2.79238462

Number of obs	=	26
F(3, 22)	=	2.94
Prob > F	=	0.0555
R-squared	=	0.2863
Adj R-squared	=	0.1890
Root MSE	=	1.5049

The F-test for nested models

H_0: **all** new β's=0 in population

H_A: **at least one** new β is not 0 in population

Numerator of F-statistic:

$(RSS_{parent} - RSS_{extended})$/(number variables added)

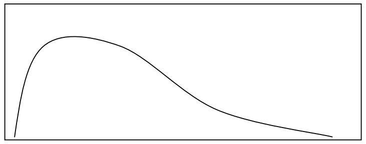
Denominator of F-statistic:

$RSS_{extended}$/(residual $df_{extended}$)

$$F_{obs} = \frac{(69.6 - 49.8)/2}{49.8/22} = 4.4$$

The F table

- Recall: the F distribution is very similar to the X^2 distribution



0

- F distribution is automatically 2-sided (like X^2)
- df change the shape of the F distribution (like X^2), but now there are two sets of df: the numerator df and the denominator df

The F table

- numerator df: # of variables added = 2
- denominator df: residual $df_{extended}$ = 22
- Using α=0.05, find F_{cr}
  - Find quantile in R, using appropriate degrees of freedom

```
> qf(.05, 2, 22, lower.tail=F)
[1] 3.443357
```

- $F_{cr}=3.44 < F_{obs}=4.4$
- So, our p-value < α

Conclusion using the F-test

- Reject H_0:
  conclude that adding year improves prediction after adjusting for housing
  - Notice:
    both individual t tests were not statistically significant, but F test was still significant
  - Must always use F test to evaluate multiple X's at once

The F test: notes

- The F test *can* be used to compare any two nested models
- If only one variable is added, it's easier to compare the models using the t test for that variable
  - t^2=F if one variable is added

The F test: how to in R

- Fit parent model
  `fit.parent <- lm(y ~ x1 + x2)`
- Fit the extended model (parent model is nested within the extended model)
  `fit.extend <- lm(y ~ x1 + x2 + x3 + x4)`
- Perform the F-test
  `print(anova(fit.parent, fit.extend))`

Example output:

```
Analysis of Variance Table

Model 1: y ~ x1 + x2
Model 2: y ~ x1 + x2 + x3 + x4
  Res.Df    RSS  Df Sum of Sq      F  Pr(>F)
1    650 110.65
2    648 109.51   2      1.14 3.3718 0.03493 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Nested Models

- Comparing nested models
  - 1 new variable: use t test for that variable
  - 2+ new variables: use F test
- Categorical predictor
  - set one group as reference
  - create dummy variable for other groups
  - include/exclude all dummy variables
  - evaluate categorical predictor with F test

Lecture 11 Summary

- Hypothesis tests in linear regression
  - Overall F-test
  - Individual coefficient 95% CI and t-tests
- F-tests for nested models
- AV plots
  - visualizing the relationship between the outcome and a continuous predictor after adjusting for the effects of a third variable