Linear Regression

Biostatistics 653

Applied Statistics III: Longitudinal Analysis

Example: Ozone Exposure Assessment

- One common problem in environmental epidemiology is determining personal exposures to environmental toxicants, such as ozone in the air. Adverse health effects associated with ozone exposure include increased incidence of cough, chest pain, and other respiratory symptoms. Although outdoor ozone concentrations are monitored by the Environmental Protection Agency (EPA), it is more difficult to determine indoor concentrations. Personal exposures, which vary based on the proportion of time spent outdoors, at home, in the workplace, and in other areas, are even more difficult to measure. Using outdoor ozone concentrations as a crude approximation of personal exposure can lead to substantial measurement error, which can in turn lead to biased parameter estimates

Study Design

- We consider data from a study conducted in State College, Pennsylvania, in which children wore small (2 cm 3 cm) personal ozone samplers. Investigators wish to model personal ozone exposures ($O_{PERSONAL}$) measured by the samplers as a function of outdoor ($O_{OUTDOOR}$) ozone concentrations (measured at a central State College site), home indoor ozone concentrations (O_{HOME}) for each child, and the proportion of time each child spent outdoors ($TIME_{OUTDOORS}$).
- The data we consider include 64 measurements of personal ozone exposure (in parts per billion or ppb) along with the corresponding measurements of outdoor ozone concentrations, home indoor ozone concentrations, and the proportion of time spent outdoors.

Look at the Data First

- Summary statistics: sample size, number of covariates, mean, variance, missingness etc.
- Data marginal normality? Transformation? Outlier?
- Scatter plot, correlation, influential point?

Model

$$\mathbf{y}_{64\times 1} = \mathbf{X}_{64\times 4}\,\beta_{4\times 1} + \varepsilon_{64\times 1}$$

$$\begin{bmatrix} 26.29 \\ 3.30 \\ 29.28 \\ 28.55 \\ 38.28 \\ \vdots \end{bmatrix}_{64\times 1} = \begin{bmatrix} 1 & 35.88 & 22.29 & 0.57 \\ 1 & 34.37 & 22.27 & 0.17 \\ 1 & 45.96 & 23.40 & 0.00 \\ 1 & 92.56 & 7.14 & 0.26 \\ 1 & 30.44 & 35.38 & 0.69 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}_{64\times 4} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}_{4\times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \vdots \end{bmatrix}_{64\times 1}$$

Parameter Interpretation

- β_0 is the intercept, which is the expected value of $O_{PERSONAL}$ when all other predictors ($O_{OUTDOOR}$, O_{HOME}, $TIME_{OUTDOORS}$) take the value zero.
- β_1 is the slope for outdoor ozone. It is interpreted as the expected ppb increase in personal exposure for a one ppb increase in outdoor ozone concentration (when all other predictors remain the same).
- β_2 is the slope for home indoor ozone. It is interpreted as the expected ppb increase in personal exposure for a one ppb increase in home indoor ozone concentration (when all other predictors remain the same).
- β_3 is the slope for the proportion of time spent outdoors. $0.01\beta_3$ is interpreted as the expected ppb increase in personal exposure for an additional one percent of time spent outdoors (when all other predictors remain the same).

SAS Code

```
proc reg data=ozone;
model personal=outdoor home timeout;
run;
```

Results

Variable	*DF*	*Estimate*	*SE*	*t value*	*p-value*
Intercept	1	3.78349	4.34206	0.87	0.3870
outdoor	1	0.09142	0.09042	1.01	0.3160
home	1	0.59544	0.16478	3.61	0.0006
timeout	1	13.64454	7.70973	1.77	0.0818

Interpretation of Results

- Intercept: the intercept is the expected response when all covariates take the value 0. For the coding scheme used for this application, the intercept is not particularly meaningful, as it is the expected personal exposure level of a child who spends no playtime outdoors and whose home and outdoor ozone exposures are 0ppm. (These values are outside the range of the observed data.)

Interpretation of Results

- A one ppb increase in outdoor ozone level is associated with a 0.09 ppb increase in expected personal exposure. However, this effect is not statistically significant.
- A one ppb increase in home ozone level is significantly associated with a 0.60 ppb increase in expected personal exposure ($p<0.01$).
- A child whose proportion of time spent outdoors is 0.10 higher has an expected personal exposure that is 1.36 ppb higher, though this association is of marginal statistical significance ($p=0.08$).

Overall Interpretation

- Personal ozone exposures appear to be closely related to indoor home ozone levels, though not to outdoor ozone levels or the amount of time spent outdoors. In particular, each ppb increase in home ozone level is associated with a 0.60 ppb increase in expected personal exposure ($p<0.01$).

Prediction

- Predicted values: a child who spends 30% of her playtime outdoors with an outdoor ozone concentration of 40 ppb and an indoor concentration of 20 ppb would be expected to have a personal exposure of $3.28 + 0.09 \times 40 + 0.60 \times 20 + 13.64 \times 0.30 = 23.47$ ppb.

Choice of L

- Time spent outdoors and home concentration are not important predictors of personal ozone exposure

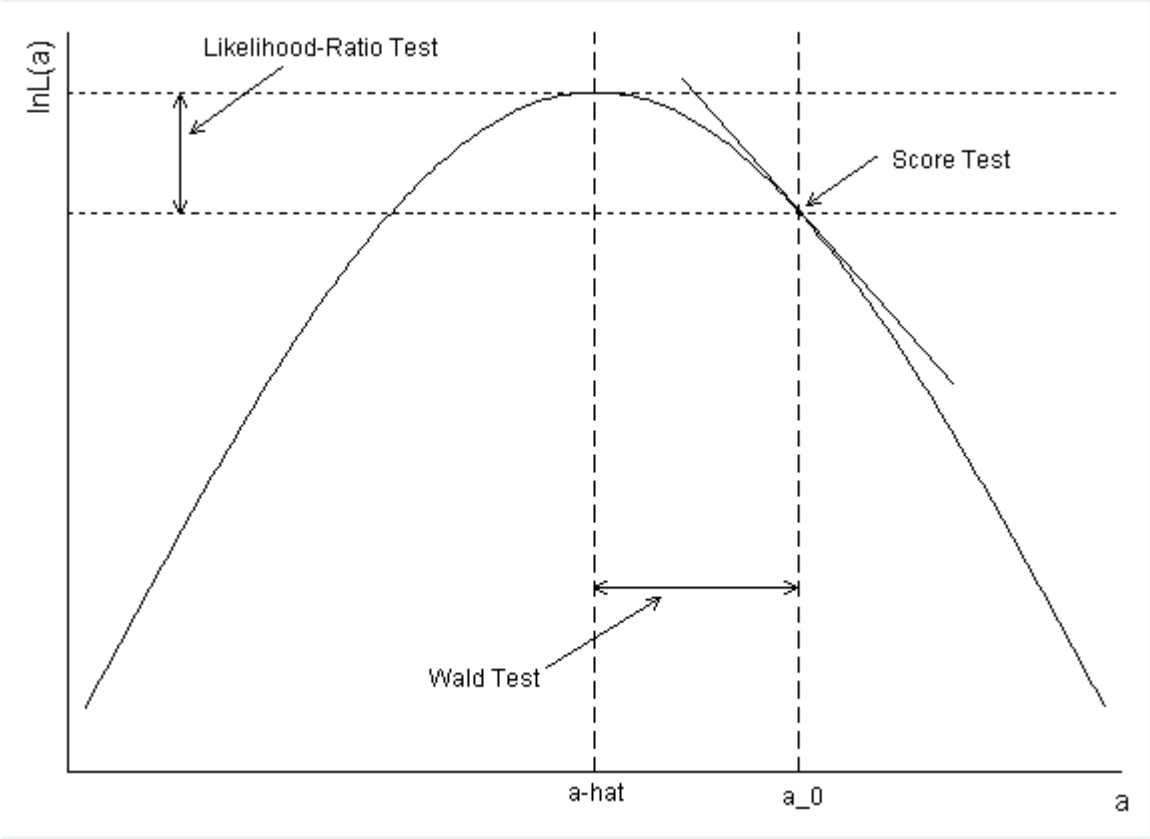
Choice of L

- A one ppb increase in outdoor ozone concentration is associated with a 0.5 ppb increase in personal exposure, and a one ppb increase in home ozone concentration is associated with a one ppb increase in personal exposure.

Choice of L

- Outdoor ozone concentration has a stronger association with personal exposure for those children who spend a greater amount of their time outdoors.

Three Tests



from UCLA

Coding Schemes

- While *reference cell* coding is the most popular coding scheme in use, other coding schemes may be convenient depending on hypotheses of interest
- We will review *reference cell*, *cell mean*, and *intercept and slope* coding schemes in the linear regression model.
- We consider the ozone data (with personal exposure as the outcome y_i) and create a new exposure variable x_i that takes value 1 if outdoor ozone is above 40 ppb and takes value 0 otherwise.

Reference Cell Coding

$$E(y_i) = \beta_0 + \beta_1 x_i$$

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \end{bmatrix}_{64 \times 4}$$

- $E(y_i | x_i = 0) = \beta_0$
- $E(y_i | x_i = 1) = L\beta = (1 \;\; 1)\beta = \beta_0 + \beta_1$

SAS Code

```
proc glm data=new;
model personal=x/solution;
estimate 'Avg Pers Exp for Low Outdoor O2' intercept 1 x 0;
estimate 'Avg Pers Exp for Low Outdoor O2' intercept 1 x 1;
run;
```

Cell Mean Coding

$$E(y_i) = \alpha_1 + \alpha_2(1 - x_i)$$

$$X = \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \end{bmatrix}_{64\times 4}$$

- $E(y_i|x_i = 0) = \alpha_2$
- $E(y_i|x_i = 1) = \alpha_1$

SAS Code

```
data new; set ozone;
ozonelow=1-x;
run;
proc glm data=new;
model personal=x ozonelow/noint solution;
run;
```

Reference Cell with Interactions

- Now we add a term, z_i, representing home ozone concentrations, to the model. We wish to test whether home ozone has the same effect on personal exposure for children who have high and low outdoor exposures, so we need an interaction term.

Reference Cell with Interactions

$$E(y_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i$$

$$X = \begin{bmatrix} 1 & 0 & 22.29 & 0 \\ 1 & 0 & 22.27 & 0 \\ 1 & 1 & 23.40 & 23.40 \\ 1 & 1 & 7.14 & 7.14 \\ 1 & 0 & 35.38 & 0 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}_{64\times 4}$$

SAS Code

```
proc glm data=ozone;
model y=x z x*z/solution;
contrast 'Intercept for Low Outdoor O2' intercept 1;
contrast 'Intercept for High Outdoor O2' intercept 1 x 1;
contrast 'Slope for Low Outdoor O2' z 1;
contrast 'Slope for High Outdoor O2' z 1 x*z 1;
run;
```

Parameter	Estimate	SE	t Value	$Pr > \|t\|$
ozonelow	10.18910496	4.67769205	2.18	0.0333
x	13.30690021	4.77004755	2.79	0.0071
homelow	0.48167987	0.25880641	1.86	0.0676
homehigh	0.65268545	0.17798005	3.67	0.0005

Intercepts and Slopes Coding

$$E(y_i) = \alpha_1 x_i + \alpha_2(1 - x_i) + \alpha_3 x_i z_i + \alpha_4(1 - x_i)z_i$$

$$X = \begin{bmatrix} 0 & 1 & 0 & 22.29 \\ 0 & 1 & 0 & 22.27 \\ 1 & 0 & 23.40 & 0 \\ 1 & 0 & 7.14 & 0 \\ 0 & 1 & 0 & 35.38 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}_{64\times 4}$$

SAS Code

Define variables

*homelow=ozonelow*z;*

*homehigh=x*z;*

And use the model statement

model y=ozonelow x homelow homehigh/noint solution;