Controlling Bias & Confounding

Chihaya Koriyama

August 5th, 2015

QUESTIONS FOR BIAS

Key concepts

- **Bias**

→ Should be minimized at the designing stage.

- **Random error**

→ Is the nature of (inaccurate) measurement.

- **Non-differential misclassification**

→ Is the nature of (inaccurate) measurement.

- **Confounding**

→ Indicative of true association. Can be controlled at the designing or analysis stage.

We can do nothing at the analysis stage!

Is the following study acceptable?

- We want to compare the mean of blood pressure levels between two groups.
- The blood pressure checker has a problem and always gives 5mmHg-higher than true values.
- All subjects were examined by the same blood pressure checker.

Internal comparison

Proper comparison between groups :

1) Comparison using accurate data

2) Comparison using (in)accurate data

As long as the magnitude of random error and bias o ong groups.

What would be the problem in this study?

FOR DISCRETE VARIABLES, MEASUREMENTS ERROR IS CALLED CLASSIFICATION ERROR OR MISCLASSIFICATION

Two types of misclassification

- **Non-differential** misclassification
  - Misclassification of a study variable that is independent of other study variables
  - Systematic error may not be a critical issue as long as it occurs in all comparison groups.
- **Differential** misclassification
  - If the error occurs in one group due to bias, then the estimate will differ from null.

Non-differential Misclassification with Two Exposure Categories

Correct Data	Exposed	Unexposed
Cases	240	200
Controls	240	600

OR = 3.0

20% of exposed subjects were misclassified

Sensitivity = 0.8
Specificity = 1.0

	Exposed	Unexposed
Cases	192	248
Controls	192	648

OR = 2.61

What is the number of each cell? Please calculate OR.

Sensitivity = 0.8
Specificity = 0.8

	Exposed	Unexposed
Cases		
Controls		

OR =

Sensitivity = 0.4
Specificity = 0.6

	Exposed	Unexposed
Cases		
Controls		

OR =

Non-differential Misclassification with Two Exposure Categories

Sensitivity = 0.8
Specificity = 0.8

	Exposed	Unexposed
Cases	232	208
Controls	312	528

OR = 1.89

20% of exposed subjects were misclassified

20% of unexposed subjects were misclassified

Sensitivity = 0.4
Specificity = 0.6

	Exposed	Unexposed
Cases	176	264
Controls	336	504

OR = 1.00

60% of exposed subjects were misclassified

40% of unexposed subjects were misclassified

How do you solve the problem of non-differential misclassification?

BIAS IN EPIDEMIOLOGIC STUDY

Different types of bias

- **Selection** bias:
  It occurs at sampling
- **Detection** bias:
  It occurs at diagnosis (outcome)
- **Measurement (information)** bias:
  It occurs at surveillance
  - **Recall** bias
  - Family information bias

You need to plan your study design carefully.

You suspect that exposure to electromagnetic field (EMF) increases the risk of childhood leukemia. And, you conducted a case-control study.

- If parents of **cases** with leukemia, **living in the neighborhood of power lines**, suspect the association and **tend to agree on participation** to the study,

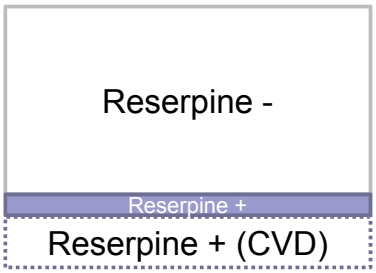
  ⇒ the association may become [ ] than what it should be.

What is this bias? How do you solve it?

In a hospital-based case-control study, the researchers excluded subjects with CVD, whom "Reserpine" was likely to be prescribed, from control group.

They found that "Reserpine was a significant risk factor of breast cancer".

Cases: Breast cancer patients



Controls: Patients at the same hospital

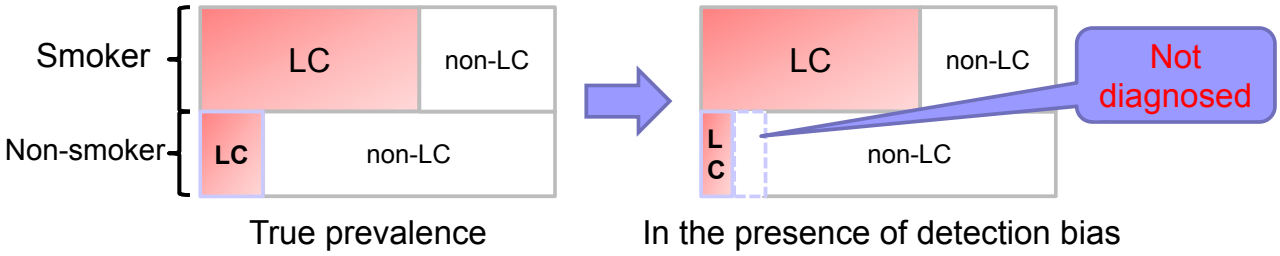
Do you agree with their conclusion?

- If you agree, why do you think so?

- If you don't agree, why do you think so? How do you solve it?

A doctor may examine the patient's chest X-ray more carefully if he knew the patient is a heavy smoker but not for non-smoking patients.

the association may become [ ] than what it should be.



What is this bias?

Detection bias

How do you avoid detection bias?

Suppose, you conducted a case-control study on relationship of prenatal infections and congenital malformations.

You asked mothers regarding prenatal episode of infections by interview / questionnaire.

Cases
(mothers of babies with defect)

Controls
(mothers of healthy babies)

What is the possible bias?

Recall bias

How do you avoid /minimize the bias?

Consider using a hospital control

Controlling for misclassification

- - **Blinding**
  - prevents investigators and interviewers from knowing case/control or exposed/non-exposed status of a given participant
- - **Form of survey**
  - mail may impose less "white coat tension" than a phone or face-to-face interview
- - **Questionnaire**
  - use multiple questions that ask same information
- - **Accuracy**
  - Multiple checks in medical records & gathering diagnosis data from multiple sources

Lecture note of Dr. Dorak (http://www.dorak.info/epi)

CONFOUNDING

3 conditions of Confounding

1. Confounders are **risk factors** for the outcome.
2. Confounders are **related to exposure** of your interest.
3. Confounders are **NOT on the causal pathway** between the exposure and the outcome of your interest.

How can we solve the problem of confounding?

"Prevention" at study design

- ✓**Limitation**
- ✓**Randomization** in an intervention study
- ✓**Matching** in a cohort study

Notice: Matching does not always prevent the confounding effect in a case-control study.

How can we solve the problem of confounding?

"Treatment " at statistical analysis

- ✓Stratification by a confounder
- ✓Multivariable / multiple analysis

Mantel-Haenszel odds ratio

- Stratification by confounding factor
  - After stratification by confounding factor, common OR, OR_{MH}, among all strata should be calculated.
  - Assumption: there is a common OR among all strata → there is no significant difference in ORs among all strata by homogeneity test.

An example of Mantel-Haenszel estimation 1

Calculate the common OR among all strata

smoking	Case	Control	
+	a_i	b_i	M_{1i}
-	c_i	d_i	M_{0i}
Total	N_{1i}	N_{0i}	T_i

$OR_c = \Sigma W_i OR_i / \Sigma w_i$

i :"i" th stratum、W_i :weight of "i" th stratum

Practice 1
Mantel-Haenszel odds ratio(1)

1. Open the "tsunagi_v1" data by excel
   - Please refer Appendix1 for the explanation of each variable.

2. Import this data set by your statistical software (STATA, R, and SPSS …)

Mantel-Haenszel odds ratio(2)

3. *Suppose, you want to examine the cancer risk by habitual alcohol drinking.*

- Please **create a contingency table** of cancer and alcohol drinking.

  STATA command: **tab alc cancer, row**

- Please **calculate an odds ratio**.

  STATA command: **cc cancer alc**
  or **cs cancer alc, or**

  Same OR but 95%CI is slightly different

Case-control study

```
. cc cancer alc
                                                         Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |        79          77  |        156       0.5064
        Controls |       316         738  |       1054       0.2998
-----------------+------------------------+------------------------
           Total |       395         815  |       1210       0.3264
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         2.396104       |    1.678901    3.416628 (exact)
 Attr. frac. ex. |         .5826558       |    .4043721    .7073138 (exact)
 Attr. frac. pop |         .2950629       |
                 +-------------------------------------------------
                               chi2(1) =    26.38  Pr>chi2 = 0.0000
```

Cohort study

```
. cs cancer alc, or
                 | alc                    |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        79          77  |        156
        Noncases |       316         738  |       1054
-----------------+------------------------+------------
           Total |       395         815  |       1210
                 |                        |
            Risk |        .2     .0944785 |   .1289256
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |         .1055215       |    .0612577    .1497852
      Risk ratio |         2.116883       |    1.584051    2.828946
 Attr. frac. ex. |         .5276074       |    .3687071    .6465115
 Attr. frac. pop |         .2671858       |
      Odds ratio |         2.396104       |    1.706524    3.364381 (Cornfield)
                 +-------------------------------------------------
                               chi2(1) =    26.38  Pr>chi2 = 0.0000
```

Mantel-Haenszel odds ratio(3)

4. Since we know that cancer risk increases with age, you may want to confirm the association between alcohol drinking and cancer risk by age group (<60, 60-69, ≥ 70).

- Please *create contingency tables* of cancer

  STATA : by age_gp, sort: tab alc cancer, row

- Please *calculate odds ratios for each age group*.

An example of Mantel-Haenszel estimation 1

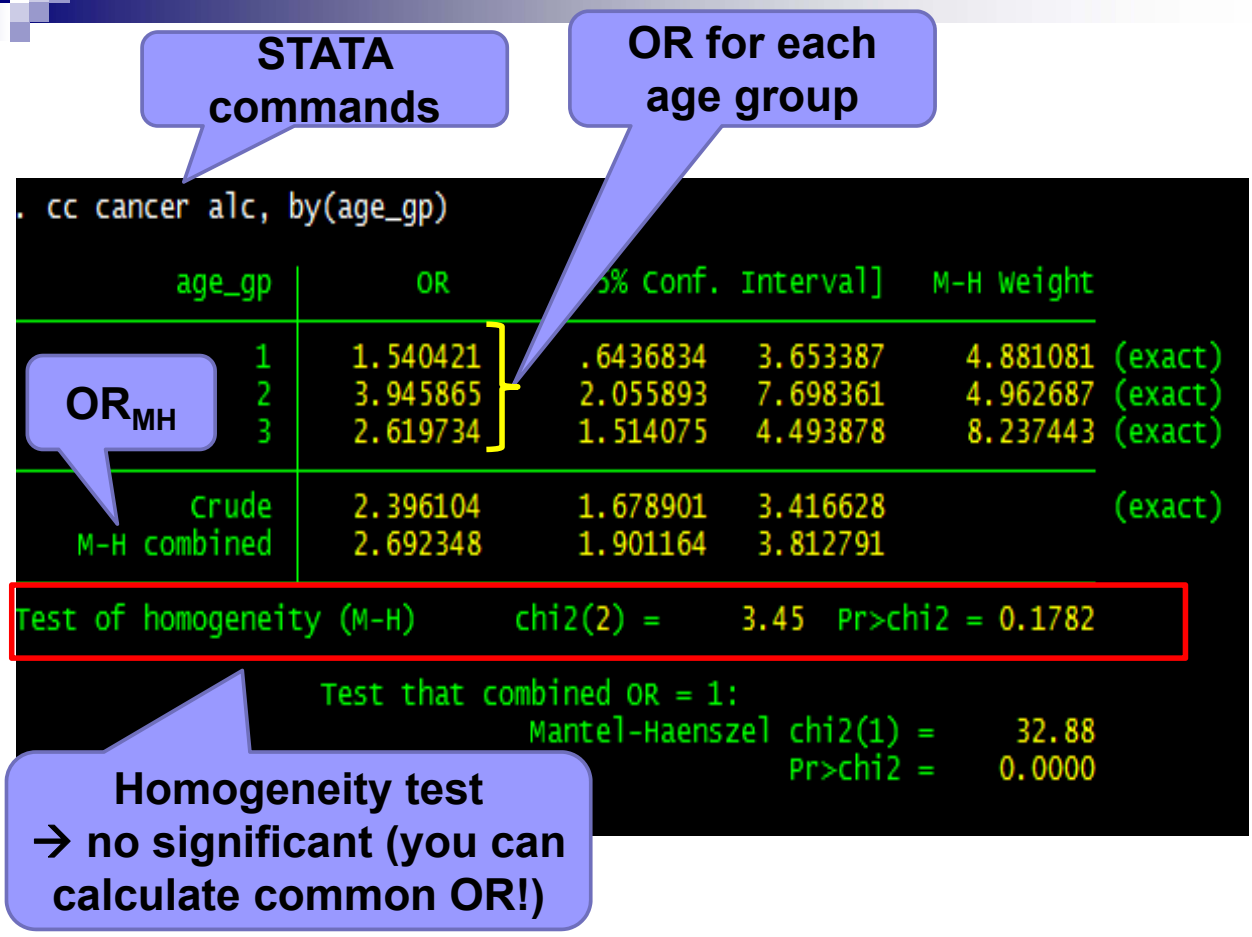
	age	alcohol	Case	Control	OR
1	<60	+	13	129	1.54
		-	14	214	1 (ref)
2	60-69	+	32	105	3.95
		-	19	246	1 (ref)
3	≥70	+	34	82	2.62
		-	44	278	1 (ref)
Total		+	79	316	2.40
		-	77	738	1

Mantel-Haenszel odds ratio(4)

5. Is there significant difference in the odds ratio among age groups?

6. Mantel-Haenszel test: homogeneity test

STATA : **cc cancer alc, by(age_gp)**



Mantel-Haenszel odds ratio(5)

You can also calculate OR_{MH} by yourself.

$$OR_{MH} = \Sigma(a_i * d_i / T_i) / \Sigma(b_i * c_i / T_i)$$

$$OR_{MH} = \frac{(13*214/370) + (32*246/402) + (34*278/438)}{(129*14/370) + (105*19/402) + (82*44/438)}$$

$$= 2.69$$

Practice 2

1. Using the "tsunagi_v1" data set, please examine the association between habitual alcohol drinking and cancer risk by sex stratification.

```
. cc cancer alc, by(male)

        male |      OR    [95% Conf. Interval]    M-H Weight
-------------+------------------------------------------------
           0 |  .9455128    .3977241    2.01076     7.399209 (exact)
           1 |  1.992308    1.179265    3.441301    11.52993 (exact)
-------------+------------------------------------------------
       Crude |  2.396104    1.678901    3.416628             (exact)
  M-H combined |  1.583126    1.061639    2.360773
-------------+------------------------------------------------
Test of homogeneity (M-H)       chi2(1) =     2.68  Pr>chi2 = 0.1014

                    Test that combined OR = 1:
                            Mantel-Haenszel chi2(1) =     4.86
                                                Pr>chi2 =    0.0276
```

Q1. Is this OR_{MH} statistically significant?

Q2. Is it OK to report OR_{MH} when the homogeneity test is statistically significant?

How can we solve the problem of confounding?

"Treatment " at statistical analysis

- ✓Stratification by a confounder
- ✓Multivariable / multiple analysis

Multivariate ≠ Multivariable (Multiple)

Am J Public Health. 2013 January; 103(1): 39–40.
Published online 2013 January. doi: 10.2105/AJPH.2012.300897
PMCID: PMC3518362
NIHMSID: NIHMS514677

Am J Public Health

Multivariate or Multivariable Regression?

Bertha Hidalgo, PhD, MPH and Melody Goodman, PhD, MS

Author information ▶ Article notes ▶ Copyright and License information ▶

See letter "Hidalgo and Goodman Respond" in volume 10 on page e1.

This article has been cited by other articles in PMC.

Abstract Go to:

The terms multivariate and multivariable are often used interchangeably in the public health literature. However, these terms actually represent 2 very distinct types of analyses. We define the 2 types of analysis and assess the prevalence of use of the statistical term multivariate in a 1-year span of articles published in the American Journal of Public Health. Our goal is to make a clear distinction and to identify the nuances that make these types of analyses so distinct from one another.

Multivariable (Multiple) analysis

A multivariable model can be thought of as a model in which multiple variables are found on the right side of the model equation. This type of statistical model can be used to attempt to assess the relationship between a number of variables; one can assess independent relationships while adjusting for potential confounders.

This is the model to control the effects of confounders!

By contrast, a multivariable or multiple linear regression model would take the form

$$(2)\ y = \alpha + x_1\beta_1 + x_2\beta_2 + \ldots + x_k\beta_k + \varepsilon$$

where y is a continuous dependent variable, x is a single predictor in the simple regression model, and x_1, x_2, …, x_k are the predictors in the multivariable model.

As is the case with linear models, logistic and proportional hazards regression models can be simple or multivariable. Each of these model structures has a single outcome variable and 1 or more independent or predictor variables.

Multivariate analysis

Multivariate, by contrast, refers to the modeling of data that are often derived from longitudinal studies, wherein an outcome is measured for the same individual at multiple time points (repeated measures), or the modeling of nested/clustered data, wherein there are multiple individuals in each cluster. A multivariate linear regression model would have the form

$$(3)\ Y_{n\times p} = X_{n\times(k+1)}\,\beta_{(k+1)\times p} + \varepsilon$$

where the relationships between multiple dependent variables (i.e., *Y*s)—measures of multiple outcomes—and a single set of predictor variables (i.e., *X*s) are assessed.

This model is to analyze the relationship between "multiple outcomes" and a single set of predictors.

LOGISTIC REGRESSION ANALYSIS

Practice 3
Multivariable analysis

1. Let's see the association between habitual alcohol drinking and cancer risk by logistic regression model.

STATA : **logistic cancer alc**
or **logit cancer alc, or**

2. Please examine this association adjusting for the effects of age and sex.

STATA : **logistic cancer alc male age**

```
. logistic cancer alc male age

Logistic regression                               Number of obs   =       1210
                                                  LR chi2(3)      =      67.45
                                                  Prob > chi2     =     0.0000
Log likelihood = -431.32695                       Pseudo R2       =     0.0725

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         alc |   1.877452   .3864541     3.06   0.002     1.254174    2.810476
        male |   2.099375   .4306218     3.62   0.000     1.404405    3.138251
         age |   1.041058   .0093757     4.47   0.000     1.022844    1.059597
------------------------------------------------------------------------------
```

STATA : logistic cancer alc male age_gp

```
. logistic cancer alc male age_gp

Logistic regression                               Number of obs   =       1210
                                                  LR chi2(3)      =      66.37
                                                  Prob > chi2     =     0.0000
Log likelihood = -431.86621                       Pseudo R2       =     0.0714

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         alc |   1.825007   .3736455     2.94   0.003     1.221779    2.726067
        male |   2.198312   .4481714     3.86   0.000     1.474193    3.278117
      age_gp |   1.669985   .1951521     4.39   0.000     1.328136    2.099824
------------------------------------------------------------------------------
```

STATA : xi: logistic cancer alc male i.age_gp

Categorical variable (>2 categories)

```
. xi: logistic cancer alc male i.age_gp
i.age_gp          _Iage_gp_1-3        (naturally coded; _Iage_gp_1 omitted)

Logistic regression                               Number of obs   =       1210
                                                  LR chi2(4)      =      66.47
                                                  Prob > chi2     =     0.0000
Log likelihood = -431.81689                       Pseudo R2       =     0.0715

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         alc |   1.825179   .3735059     2.94   0.003     1.222123    2.725812
        male |   2.192979   .4473008     3.85   0.000     1.470332    3.270798
  _Iage_gp_2 |   1.792761   .4573496     2.29   0.022     1.087359    2.955776
  _Iage_gp_3 |    2.84385   .6937492     4.28   0.000     1.763025    4.587276
------------------------------------------------------------------------------
```

If there is no linear trend of the cancer risk by age,
it would be better to use categorical variable for age.

REGRESSION ANALYSIS

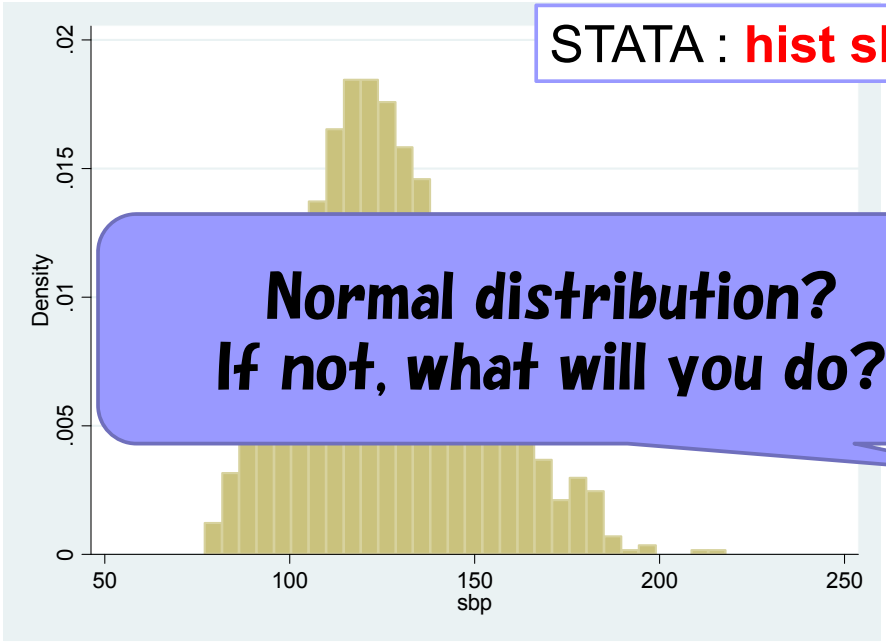
Practice 4
Regression analysis (1)

- Suppose, you want to know predictors of systolic blood pressure in the subjects of "tsunagi_v1" data.

- What do you have to check first?

Distribution of systolic blood pressure

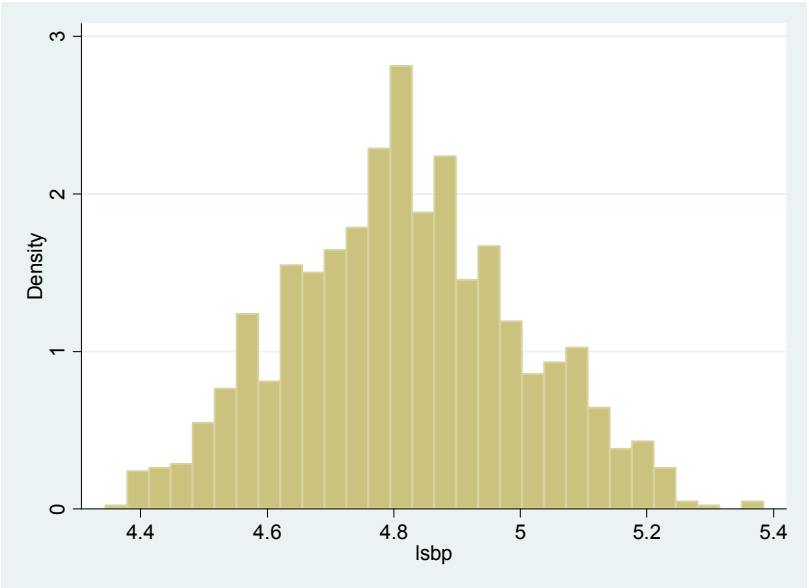
STATA : **hist sbp**



Normal distribution?
If not, what will you do?

Log-transformation may work…

STATA : **gene lsbp=log(sbp)**
hist lsbp



Practice 4
Regression analysis (2)

- Age is one of the predictors of systolic blood pressure.

- Please conduct regression analysis using "age" as a explanatory variable.

STATA : **reg lsbp age**

STATA commands

```
. reg lsbp age
```

Source	SS	df	MS	Number of obs =	1210
				F(1, 1208) =	91.90
Model	2.84842138	1	2.84842138	Prob > F =	0.0000
Residual	37.4400935	1208	.030993455	R-squared =	0.0707
				Adj R-squared =	0.0699
Total	40.2885149	1209	.033323834	Root MSE =	.17605

lsbp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0044274	.0004618	9.59	0.000	.0035213	.0053334
_cons	4.531922	.0301251	150.44	0.000	4.472819	4.591026

SBP= 4.531922 + **0.0044274***age

This indicates that SBP will increase 0.004 per age (year).

Practice 4
Regression analysis (3)

- Please transform age variable into 10-year age group.

STATA : gene age10=floor(age/10)

- Let's see the association between age and systolic blood pressure using this variable (age10).
- What do you expect?

```
. reg lsbp age10
```

Source	SS	df	MS
Model	2.8196114	1	2.8196114
Residual	37.4689035	1208	.031017304
Total	40.2885149	1209	.033323834

Number of obs	1210
F(1, 1208)	90.90
Prob > F	0.0000
R-squared	0.0700
Adj R-squared	0.0692
Root MSE	.17612

lsbp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age10	.0431853	.0045294	9.53	0.000	.0342989	.0520717
_cons	4.557933	.0276	165.14	0.000	4.503784	4.612083

SBP= 4.557933 + **0.0431853***age(10)

cf. SBP= 4.531922 + **0.0044274***age

Practice 4
Regression analysis (4)

- Suppose, hemoglobin level may be one of the predictors of systolic blood pressure.
- Please pick-up other potential predictors (other than hemoglobin) for systolic blood pressure in this data set based on your knowledge.
- And, conduct regression analysis.

How many explanatory variables can we use in a model?

Model	Number of explanatory variables	Example
Linear regression model	**Sample size / 15**	Up to around 6-7 variables in **100 subjects**
Logistic regression model	**Smaller sample size of outcome / 10**	Up to 10 variables if the numbers of cases and controls are **100** and 300, respectively.
Cox proportional hazard model	**The number of event / 10**	Up to 9 variables if you have **90 events** out of 150 subjects

ATTENTION!

- When you include categorical variable in your model, you have to count that variable as (the number of categories – 1).
  - For example, the variable of age group used in the previous practice, we have to count it as "two" (=3 categories – 1) variables.