Lecture Notes:
Linear Regression
DrPH Multivariable Course

Kamarul Imran Musa
Professor (Epidemiology and Statistics)

2023-11-14

Contents

1 Linear Regression: A Lecture Note

1.1 Introduction

At the end of the lecture, you should be able to:

1. understand the concepts for multivariable statistical analysis in public health
2. describe the approaches for multivariable statistical analysis in public health
3. to analyze and interpret linear regression from statistical software
4. be able to appraise the common multivariable analysis in the medical and public health literature

Specifically, we will learn about

1. model building in linear regression
2. role of variables in the multivariable model: either as a confounder, a mediator or a moderator (an effect-modifier)
3. interpreting parameters estimated from linear regression

4. checking and assessment for linear regression models
5. remedial process for linear models when assumptions are not met
6. making predictions from linear models

1.2 Motivation

Linear regression is useful when we want to predict the average response of a continuous outcome variable or a continuous dependent variable as a function of one covariate or a set of covariates. In linear regression we need to assume that the outcome or dependent variable follows Gaussian or normal distribution as a function of the covariate or covariates.

1.2.1 Motivating example

Linear regression is suitable to analyze the relationship of an assumed normally distributed outcome as a function of a covariate (or independent variable) or a set of covariates (more than one independent variables).

Because the distribution of the outcome variable needs to be normally distributed given the covariate or covariates, it is suitable for many data with the numerical or continuous outcome.

1.3 Models

1.3.1 Simple linear regression

In simple linear regression, the model has two variables.

1. One is the dependent or outcome variable. This variable numerical or continuous in nature assumed to be normally distributed given the covariate or covariates.
2. One independent variables. The independent variable can be a numerical variable or can be a categorical variable.

The dependent variable is also known as the outcome variable. And the independent variable can be called as the predictor, the regressor or the covariate. The distribution of the outcome variable should be normal or Gaussian as a function of the covariate.

The sample is the object of our research. It can be a grouped of people, patients, animals or laboratory tests. Using the estimated parameters we can make inference of the results to the population of interest. Based on the parameters, we develop the understanding of the relationship between the two variables - the dependent and the independent - in which would be useful in daily life such as in clinical settings or in public health settings.

The name *regression* indicates the use of regression model in our data analysis. One of the product of regression models is an equation known as the *linear equation*. This equation is important. To be able to understand how regression works and means, it is imperative to understand the linear equation. Linear equation is then used to predict the value of a dependent variable given the value of an associated independent variable.

1.3.2 Multiple linear regression

Intuition tells us that, in general, we ought to be able to improve our predicting ability by including more independent variables in such an equation. The concepts and techniques for analyzing the associations among several variables are natural extensions of those explored in the previous topics such as simple linear regression.

The response of an experiment or observation depends not only on a single predictor. But the more predictors you have, the more complicated the model becomes. Still the model has to assume its Gaussian distribution in its dependent variable as a function of more than one covariates.

Hence, in multiple linear regression, the model has three or more variables.

1. One is the dependent or outcome variable. This variable is numerical or continuous in nature with normally distribution given the covariate or covariates.
2. Two or more independent variables. The independent variables can be all numerical, or all categorical or a mixed of numerical and categorical variables.

1.4 Statistical concepts

In multiple regression model, we assume that a linear relationship exists between some variable Y, which we call the dependent variable (or the outcome or the regressand), and k independent variables (or the predictor, covariate, explanatory or the regressor) such as $X_1, X_2, ..., X_k$.

The independent variables are sometimes referred to as explanatory variables, because of their use in explaining the variation in Y. They are also called predictor variables, because of their use in predicting Y and covariates.

1.5 Model assumptions

The assumptions in simple linear regression still apply in multiple linear regression. As in simple linear regression, we test the assumptions after running the analysis (most of us will perform this using a statistical software).

The assumptions for multiple linear regression are:

1. The X_i are non-random (fixed) variables. This condition indicates that any inferences that are drawn from sample data apply only to the set of X values observed and not to some larger collection of X_s.

2. For each set of X_i values there is a subpopulation of Y values which are normally distributed. To construct certain confidence intervals and test hypotheses, it must be known, or the researcher must be willing to assume, that these subpopulations of Y values are normally distributed.
3. The variances of the sub-populations of Y are all equal.
4. The Y values are independent. That is, the values of Y selected for one set of X values do not depend on the values of Y selected at another set of X values.

Figure @ref(fig:LinearityAssumption) sums the first 3 assumptions:

```
## here() starts at C:/Users/drkim/OneDrive - Universiti Sains Malaysia/3_Statistics/Linear_Regression_
```

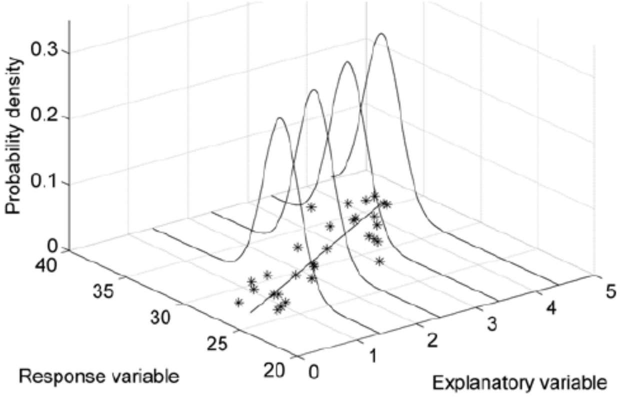


Figure 1: Linearity Assumptions

You can read more here

1.6 Model and equation

1.6.1 General linear models

Generally, the multiple linear regression model is written as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \epsilon_i$$

This form is also known as the first-order model. The Y_i is the typical value from one of the subpopulation. The β_i is called as the regression coefficient. The $x_{1j}, x_{2j}, ..., x_{kj}$ are the particular values for $X_{1j}, X_{2j}, ..., X_{kj}$. The ϵ_i is a random error variable or error term. Its mean equals 0 and variance of σ^2. The ϵ_i are independent with $N(0, \sigma^2)$

1.6.2 Linear regression function

The regression function for a multiple linear model with predictors X_1 and X_2 is written as an equation such as: $E\left\{Y_i\right\} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki}$.

The random error variable, which is an error term (or noise), is calculated by $\epsilon_i = Y_i - E\left\{Y_i\right\}$.

1.6.3 Polynomial regression models

A polynomial regression model produces a curvilinear response variables. In polynomial regression, each predictor variable is represented by various powers.

The example below shows a polynomial regression model raised to the first and second powers and is known as the second order models $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2$

1.6.4 Qualitative or categorical covariates

As in simple linear regression, a qualitative (a categorical variable) covariate can be fitted in a multiple linear regression model. So in multivariable linear models (like any other multivariable models), the covariates commmoly consist of a mixed of numerical and categorical covariates.

In order to incorporate such qualitative independent variable in the multiple regression model, it must be quantified in some manner. This quantification can be accomplished through the use of what are known as *dummy variables*.

For a multiple linear regression model such as this:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \epsilon_i$$
$$Y_i = \beta_0 + \beta_1 AGE_{1i} + \beta_2 SEX_{2i} + \epsilon_i$$

$$X_2 = \begin{cases} male & X_2 = 1 \\ female & X_2 = 0 \end{cases}$$

The expected response function for female would be:

$$\widehat{Y_i} = \hat{\beta}_0 + \hat{\beta}_1 AGE_{1i} + \hat{\beta}_2 \times 0$$

and for male:

$$\widehat{Y_i} = \hat{\beta}_0 + \hat{\beta}_1 AGE_{1i} + \hat{\beta}_2 \times 1$$
$$\widehat{Y_i} = \hat{\beta}_0 + \hat{\beta}_2 + \hat{\beta}_1 AGE_{1i}$$

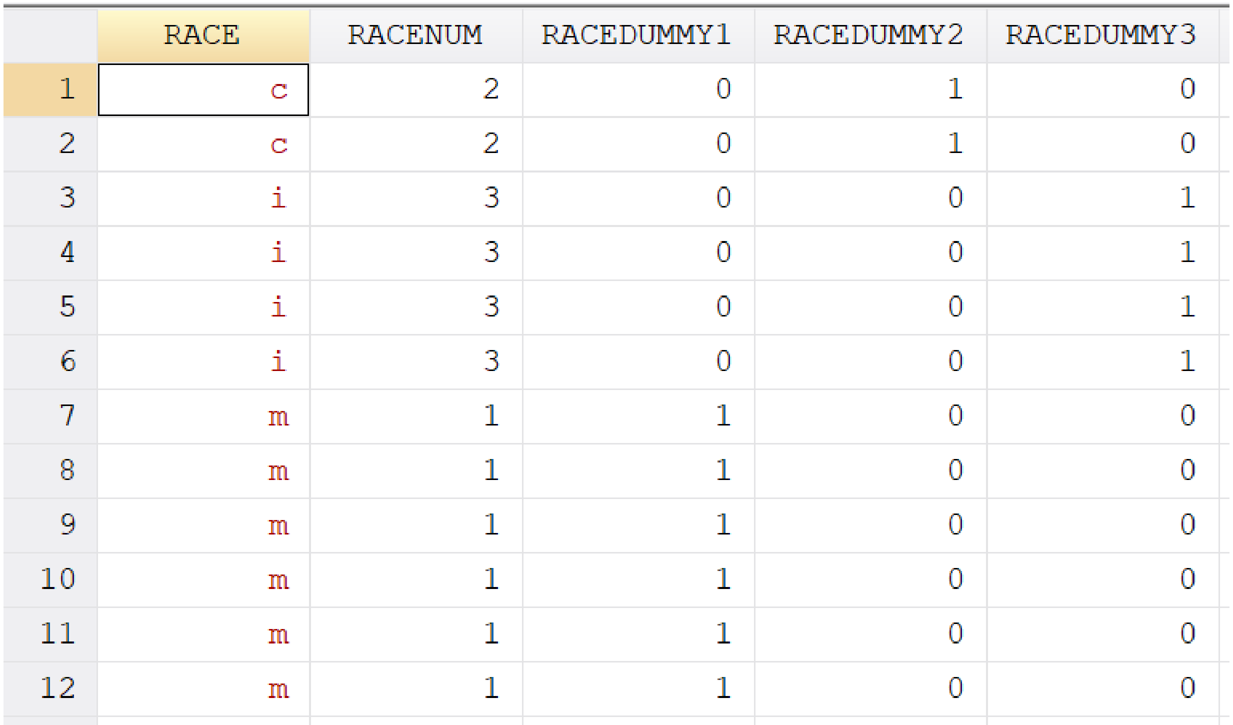
In the background, the statistical software using a certain function will create additional columns (dummy variables) to facilitate the interpretation of regression.

Figure @ref(fig:dummyvar) shows the dummy variables created in Stata.

1.6.5 Terms and parameters

We call regression coefficients as *betas* or β_s. Each β is called as a parameter. β_0 is the Y intercept of the regression plane when sum of all other product of covariates and their regression parameters equal 0. The mean response refers to $E\{Y_i\}$ at the certain values of X_i.

In linear regression, the term partial regression coefficients refer to the partial effect of one predictor variable when the other variable is included in the model and is held constant.



	RACE	RACENUM	RACEDUMMY1	RACEDUMMY2	RACEDUMMY3
1	c	2	0	1	0
2	c	2	0	1	0
3	i	3	0	0	1
4	i	3	0	0	1
5	i	3	0	0	1
6	i	3	0	0	1
7	m	1	1	0	0
8	m	1	1	0	0
9	m	1	1	0	0
10	m	1	1	0	0
11	m	1	1	0	0
12	m	1	1	0	0

Figure 2: Dummy variables

1.6.6 Estimation

There two methods that can be used to estimate the parameters in linear regression:

1. Least square estimation
2. Maximum likelihood estimation

Both methods are:

1. unbiased minimum variance
2. consistent
3. sufficient

1.6.6.1 Least square methods The unbiased estimates of the parameters $\beta_0, \beta_1, ..., \beta_k$ are obtained by the method of least squares. This means that the sum of the squared deviations of the observed values of Y from the resulting regression surface is minimized.

1.6.6.2 Maximum likelihood estimation We will not discuss this for now. You can refer to other text if you are interested to know more.

1.7 Linear regression examples

Table @ref(tab:lin-mod1) is the output of a linear regression model with hba1c as the outcome variable and fasting blood sugar (fbs) as the covariate.

Table 1: Fasting blood sugar (fbs) as the covariate (in mmol/l).

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	3.6768558	0.0384042	95.74103	0	3.6015633	3.7521484
fbs	0.3784035	0.0061756	61.27419	0	0.3662961	0.3905109

A multiple linear regression model with fasting blood sugar and modified oral glucose tolerance test (mogtt2h) as covariates is shown in Table @ref(tab:linmodak1).

Table 2: Fasting blood sugar and MOGTT at 2 hours as covariates

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	3.9399850	0.0343565	114.67956	0	3.8726254	4.0073445
fbs	0.1380130	0.0086038	16.04099	0	0.1211444	0.1548816
mogtt2h	0.1173771	0.0044443	26.41084	0	0.1086636	0.1260906

1.8 Model comparisons

We can make use of the coefficient of multiple determination to assess if two models are different. As you remember that the total variation present in the Y values may be partitioned into two components:

1. The explained variation, which measures the amount of the total variation that is explained by the fitted regression surface
2. The unexplained variation, which is that part of the total variation not explained by fitting the regression surface

The coefficient of multiple determination measures the proportionate reduction of total variation in Y associated with the use of set of X variables, that is:

$$SST = SSR + SSE$$

This means, the Total Sum of Squares (SST) = Explained Sum of Squares (SSR) + Unexplained Sum of Squares (SSE)

This can be represented as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

We can use the coefficient of multiple determination to compare between different models using *anova* function in R software for example. When comparing models, we set the null hypothesis as the two models are similar and the alternative hypothesis as the two models are different.

The level of significance of p-value of < 0.05 is used to indicate significant difference. If $p < 0.05$, the two models are different and we have to choose which one fits our modelling better.

Note that the coefficient of multiple determination or R^2 is $0 \geq R^2 \leq 1$. Adding more X variables to the model increases R^2 and never reduces it. In view of that, check the adjusted multiple determination that divides sum of squares by its associated degree of freedom

The obtain the coefficient of multiple correlation is simple. It can be derived from $R = \sqrt{R^2}$

R-squared has additional problems that the adjusted R-squared and predicted R-squared are designed to address.

1. Every time you add a predictor to a model, the R-squared increases, even if due to chance alone.
2. If a model has too many predictors and higher order polynomials, it begins to model the random noise in the data. This condition is known as overfitting.

The adjusted R-squared compares the explanatory power of regression models that contain different numbers of predictors. Suppose you compare a five-predictor model with a higher R-squared to a one-predictor model. Does the five predictor model have a higher R-squared because it's better? Or is the R-squared higher because it has more predictors?

Use the adjusted R-square to compare models with different numbers of predictors

Read more here

1.8.1 Overall F test

The overall F test assesses if there is significant regression relation between variable Y and the set of X variables.

$$H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1} = 0$$
$$H_a : \text{not all } \beta_{k,(k=1,...,p-1)} = 0$$

The test statistic is calculated using F-test:

$$F^* = \frac{MSR}{MSE}$$

1.8.2 Interaction between covariates

When the effects of the predictor variables on the response variable are not additive, the effect of one predictor variable depends on the levels of the other predictor variables. The two variables interact with each other but the effect of the interaction is not additive. The usual interpretation of regression parameters that is one unit increase in X will increase Y by certain units, can not be used.

The model where interaction (non-additive model) is present can be modelled as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \epsilon_i$$

the $X_{1i}X_{2i}$ is called as the two-way interaction term. This model is a special case of the general linear regression model. If we let $X_{3i} = X_{1i}X_{2i}$, then the equation can be written as: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$

Let us say we have a linear equation of $E\{Y_i\} = 10 + 2X_{1i} + 5X_{2i} + 0.5X_1X_2$, then note that the slopes of the response functions plotted against X, now differ for $X_2 = 1$ and $X_2 = 3$.

The slope of the response function when $X_2 = 1$ is

$$E\{Y\} = 10 + 2X_1 + 5 \times 1 + 0.5X_1 \times 1$$
$$E\{Y\} = 15 + 2.5X_1$$

but when $X_2 = 3$, it is

$$E\{Y\} = 10 + 2X_1 + 5 \times 3 + 0.5X_3 \times 1$$
$$E\{Y\} = 25 + 3.5X_1$$

1.8.2.1 Interaction between a numerical covariate with another numerical covariate

Below is the multiple linear regression model with fasting blood sugar, modified oral glucose tolerance test and the 2-way interaction between them as covariates is shown in Table @ref(tab:linmod2k)

Table 3: Fasting blood sugar, MOGTT at 2 hours and their 2-way interaction as covariates

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	4.8925568	0.0616124	79.408589	0.0000000	4.7717590	5.0133546
fbs	-0.0184767	0.0118797	-1.555324	0.1199547	-0.0417681	0.0048146
mogtt2h	0.0298382	0.0064056	4.658156	0.0000033	0.0172794	0.0423970
fbs:mogtt2h	0.0116869	0.0006391	18.286633	0.0000000	0.0104339	0.0129400

1.8.2.2 Interaction between a numerical covariate and a categorical covariate

This may be a model with fasting blood sugar, receptor status(1 = positive, 0 = negative) and the 2-way interaction between them as covariates in shown in Table @ref(tab:linmod3k)

Table 4: Fasting blood sugar (FBS), receptor status and their two-way interaction term as the covariates

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	3.6859620	0.0385479	95.620351	0.0e+00	3.6103878	3.7615363
fbs	0.3750613	0.0062455	60.053027	0.0e+00	0.3628168	0.3873058
receptorpositive	3.7841430	0.5258004	7.196919	0.0e+00	2.7532953	4.8149906
fbs:receptorpositive	-0.2019069	0.0455591	-4.431755	9.6e-06	-0.2912269	-0.1125869

1.8.2.3 Interaction between a categorical and a categorical covariate

This model may include a variable with receptor status(1 = positive, 0 = negative) , male sex (2 = yes, 1 = no) and the 2-way interaction between them as covariates in shown in Table @ref(tab:linmod3bk)

Table 5: Receptor status, sex and their two-way interaction term as the covariates

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	5.8118200	0.0368770	157.6002610	0.0000000	5.7395220	5.8841181
maleyes	-0.0421066	0.0457643	-0.9200743	0.3575860	-0.1318284	0.0476153
receptorpositive	2.9770689	0.4757621	6.2574733	0.0000000	2.0443276	3.9098101
maleyes:receptorpositive	0.8657177	0.5946770	1.4557779	0.1455276	-0.3001586	2.0315939

From the models, we need to be able to

1. write the regression equations
2. interpret the regression parameters (estimation and inference)
3. compare the models
4. make predictions

1.8.3 Confounders, mediators and moderators (effect-modifier)

Confounder and mediators are covariates. But both have similarities and differences. Both changes the coefficients. Both are (again) covariates.

Confounders are variables that cause the independent variable of interest or share a common cause with it. Mediators in contrast are consequences of the independent variable of interest.

The decision whether a covariate is a confounder or a mediator is not solely based on statistical analysis but instead (and more importantly) is based on subject matter knowledge

A linear regression model with Systolic BP as the covariate is shown in Table @ref(tab:linsbp)

Table 6: Systolic BP as covariate

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	4.656250	0.1231713	37.803050	0	4.4147703	4.8977298
sbpr1	0.008531	0.0009000	9.479169	0	0.0067666	0.0102954

A linear regression model with Diastolic BP as the covariate is shown in Table @ref(tab:lindbp)

Table 7: Distolic BP as covariate

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	4.5689736	0.1384995	32.989091	0	4.2974425	4.8405046
dbpr1	0.0156158	0.0017275	9.039788	0	0.0122291	0.0190025

A multiple linear regression model with Systolic BP and Diastolic BP as the covariates is shown in Table @ref(tab:linsbpdbp)

Table 8: Systolic and Distolic BP as covariates

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	4.4019952	0.1431268	30.755920	0.0000000	4.1213923	4.6825980
sbpr1	0.0055594	0.0012406	4.481121	0.0000076	0.0031271	0.0079916
dbpr1	0.0082673	0.0023791	3.475008	0.0005159	0.0036031	0.0129316

Examine:

1. the regression parameter for the covariate SBP in the simple (univariable analysis)
2. the regression parameter for the covariate DBP in the simple (univariable analysis)
3. the regression parameters for both the covariates SBP and DBP in the multiple (multivariable analysis)

Is there a confounding effect or a mediating effect? Which confound which and which mediate which?

Read how to deal with mediators and interaction in Applied Logistic Regression, chapter Special Topics, section Mediation and Statistical Interaction (page 441-455) (Hosmer 2013).

1.8.4 Multicollinearity

When the predictor variables are **highly correlated** among themselves, intercorrelation or multicollinearity among them is said to exist when the predictor variables are uncorrelated, the effects ascribed to them by a first-order regression model are the same no matter which other of these predictor variables are included in the model.

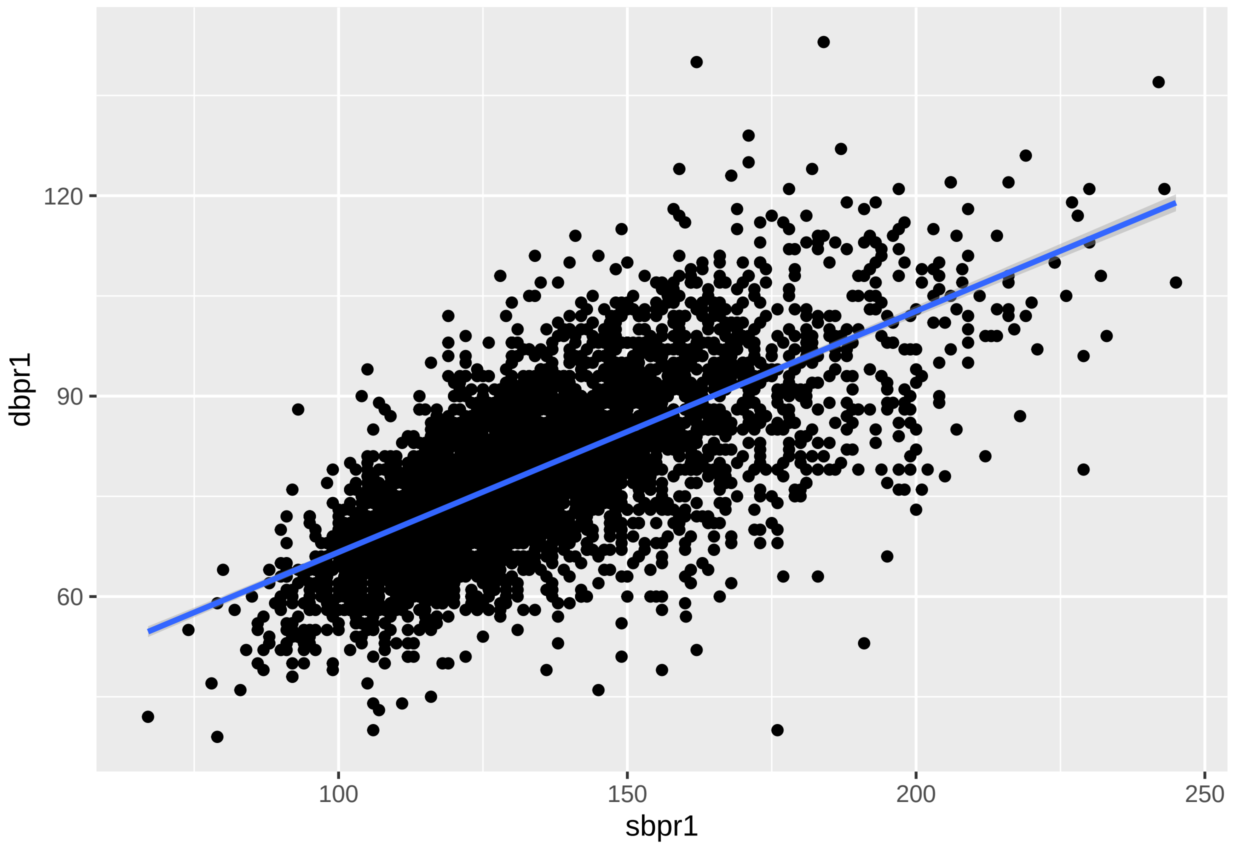


Figure 3: Scatter plot with the linear fit line between DBP and SBP

To test the possible existence of multicollinearity, pairwise coefficients of simple correlation between the predictor variables can be performed. For example the relationship between SBP and DBP is complicated by multicollinearity because the correlation between them equals 0.6854878 which is high (See Figure @ref(fig:sbpdbpcor). The issues discussed in the confounding and mediation section will resurface.

When more than two predictor or independent variables are highly correlated, the regression coefficient of anyone variable depends on which other predictor variables are included in the model and which ones are left out. Thus, a regression coefficient does not reflect any inherent effect of the particular predictor variable on the response variable but only a marginal or partial effect, given whatever other correlated predictor variables are included in the model.

The problem of multicollinearity is easy to detect if only 2 covariates are highly correlated. But it becomes complicated when more than 2 covariates are highly correlated. For example, what happen to the model with these covariates fasting blood sugar, random blood sugar, modified OGTT at 1 hour, modified OGTT at 2 hours are modelled together.

Similar problem can happen if categorical covariates are highly correlated between them. How would you check for colinearity between categorical variables?

1.9 Model checking

1.9.1 Linear regression assumptions

We can perform some model checking procedures or diagnostics procedures to check our linear model. For example, sctterplot matrix for predictor against the response variable to look for strength of association, nature, identify gaps and outlying points.

We can examine the residual plots by

1. plotting residuals against the fitted values
2. plotting residuals against predictors (inside and outside of the model)
3. running correlation test for normality
4. running Brown-Forsythe test for constance of error variance

In the plots, look for any curvature of the residuals and also for the constancy of variance.

1.9.2 Making inferences from linear regression models

We used the least square or maximum likelihood estimators for regression parameters. Both are unbiased. For the normal error regression, the confidence limits for $b_k \pm t_{(1-\alpha/2;n-p)} SE_{b_k}$.

The test for β_k:

$$H_0 : \beta k = 0$$
$$H_a : \beta_k \neq 0$$

The test statistics that can be used to conclude H_0 or H_a is:

$$t^* = \frac{b_k}{SE_{b_k}}$$

1.10 Making prediction based on linear regression models

To make prediction of the outcome variables based on the selected model, follow these steps

1. write the regression equation
2. subsitute the value of X_i
3. calculate the fitted value of the outcome variable

The predicted values (fitted values, their standard error, residuals and other values) that are based on

1. model with FBS, MOGTT at 2 hours and their interaction as the covariates is shown in Table @ref(tab:predmod21a)

Table 9: FBS, MOGTT at 2 hours and their interaction as the covariates

.rownames	hba1c	fbs	mogtt2h	.fitted	.resid	.hat
1	4.9	2.58	1.67	4.945071	-0.0450710	0.0022429
2	5.2	5.63	4.01	5.172032	0.0279680	0.0008719
5	5.3	4.84	4.74	5.212680	0.0873203	0.0004677
6	5.4	4.64	6.52	5.354933	0.0450674	0.0002941
7	5.0	3.93	6.43	5.307131	-0.3071308	0.0004423
8	5.3	5.97	8.84	5.662797	-0.3627966	0.0004438

2. model with FBS, insulin and their interaction is shown in Table @ref(tab:linmod31) and Table @ref(tab:linmod32)

Table 10: FBS, insulin and their interaction as the covariates

.rownames	hba1c	fbs	receptor	.fitted	.resid	.hat
1	4.9	2.58	neg	4.653620	0.2463799	0.0005676
2	5.2	5.63	neg	5.797557	-0.5975571	0.0002398
3	10.2	5.14	neg	5.613777	4.5862230	0.0002475
4	7.6	6.09	neg	5.970085	1.6299148	0.0002483
5	5.3	4.84	neg	5.501259	-0.2012586	0.0002607
6	5.4	4.64	neg	5.426246	-0.0262464	0.0002731

Table 11: FBS, insulin and their interaction as the covariates

.rownames	hba1c	fbs	receptor	.fitted	.resid	.hat
4335	5.4	5.15	neg	5.617528	-0.2175276	0.0002472
4336	4.4	3.75	neg	5.092442	-0.6924418	0.0003630
4337	4.8	4.14	neg	5.238716	-0.4387157	0.0003166
4338	5.4	4.57	neg	5.399992	0.0000079	0.0002781
4340	4.5	5.46	neg	5.733797	-1.2337966	0.0002406
4341	5.6	4.94	neg	5.538765	0.0612352	0.0002556

Can you calculate the average fasting blood sugar (fbs) manually for each of the models above?

1.10.1 Fitted values, outliers, leverage and influence

Outliers: observations with large residuals (the deviation of the predicted score from the actual score), note that both the red and blue lines represent the distance of the outlier from the predicted line at a particular value of enroll

Leverage: measures the extent to which the predictor differs from the mean of the predictor; the red residual has lower leverage than the blue residual

Influence: observations that have high leverage and are extreme outliers, changes coefficient estimates drastically if not included

Read more here

Standardized residuals: We are looking for values greater than 2 and less than -2 (outliers)

Leverage: a school with leverage greater than $(2k + 2)/n$ should be carefully examined. Here k is the number of predictors and n is the number of observations, so a value exceeding $(2 \times 1 + 2)/400 = 0.01$ would be worthy of further investigation. Find observation that has both a large standardized residual and leverage, which suggests that it may be influential.

Cook's Distance: Now let's look at Cook's Distance, which combines information on the residual and leverage. The lowest value that Cook's D can assume is zero, and the higher the Cook's D is, the more influential the point is. The conventional cut-off point is $4/n$, or in this case 4/400 or 0.01.

DFBETA: Cook's Distance can be thought of as a general measure of influence. You can also consider more specific measures of influence that assess how each coefficient is changed by including the observation. Imagine that you compute the regression coefficients for the regression model with a particular case excluded, then recompute the model with the case included, and you observe the change in the regression coefficients due to including that case in the model. This measure is called DFBETA and a DFBETA value can be computed for each observation for each predictor.

1.11 Remedial measures for problematic models

1.11.1 Model non-normal data

Often one may wish to attempt a transformation of the data. Mathematical transformations are useful because they do not affect the underlying relationships among variables.

Since hypothesis tests for the regression coefficients are based on normal distribution statistics, data transformations can sometimes normalize the data to the extent necessary to perform such tests.

Simple transformations,such as taking the square root of measurements or taking the logarithm of measurements, are quite common.

1.11.2 Model with unequal variance

When the variances of the error terms are not equal, we may obtain a satisfactory equation for the model, but, because the assumption that the error variances are equal is violated, wewill not be able to perform appropriate hypothesis tests on the model coefficients.

Just as was the case in overcoming the non-normality problem, transformations of the regression variables may reduce the impact of unequal error variances.

1.11.3 Model with correlated independent variables

Multicollinearity is a common problem that arises when one attempts to build a model using many independent variables. Multicollinearity occurs when there is a high degree of correlation among the independent variables. For example, imagine that we want to find an equation relating height and weight to blood pressure. A common variable that is derived from height and weight is called the body mass index (BMI).

The least complex solution to multicollinearity is to calculate correlations among all of the independent variables and to retain only those variables that are not highly correlated.

A conservative rule of thumb to remove redundancy in the data set is to eliminate variables that are related to others with a significant correlation coefficient above 0.7.

1.11.4 Box-Cox transformation

A possible transformation for Y or the covariates.

1.11.5 Centering and colinearity

The need for centering a predictor variable. Centering a numerical variable can reduce collinearity amongst covariates. This is especially true for example in polynomial regression models.

1.12 Presentation of results

- Justify the use of normal distribution regression
- Justify including or excluding covariates
- Compare and assess the models
- Are the models problematic. What are the remedial measures?
- Show the main models in the main text
- Show other models developed in the appendix
- Provide results from model assessment in the main text or as appendices

1.13 References

1.13.1 Compulsory text

The main reference is in Kutner. Applied Linear Statistical Models book. Please read chapter 6,7 and 8 (Neter 2013). To understand model building, you may want to read David W Hosmer, Stanley Lemeshow, Rodney X Sturdivant Applied Logistic Regression book on chapter 4, 10.8 and 10.9 (Hosmer 2013).

1.13.2 Additional texts

The concept of multivariable is explained in Mitchell H. Katz. Multivariable Analysis: A Practical Guide for Clinicians and Public Health Researchers book (Katz 2011). To practise using R for linear regression, you may want to read Julian J. Faraway. Linear Models with R book (Faraway 2016).

Faraway, J. J. 2016. *Linear Models with r.* Chapman & Hall/CRC Texts in Statistical Science. CRC Press. https://books.google.com.my/books?id=i0DOBQAAQBAJ.

Hosmer, David. 2013. *Applied Logistic Regression.* Hoboken, New Jersey: Wiley.

Katz, M. H. 2011. *Multivariable Analysis: A Practical Guide for Clinicians and Public Health Researchers.* Cambridge University Press. https://books.google.com.my/books?id=-X4G4dHsARQC.

Neter, John Neter Michael Kutner William Wasserman Christopher Nachtsheim John. 2013. *Applied Linear Statistical Models 5ed (Pb 2013)*. Paperback. MC GRAW HILL INDIA. https://lead.to/amazon/com/?op=bt&la=en&cu=usd&key=1259064743.