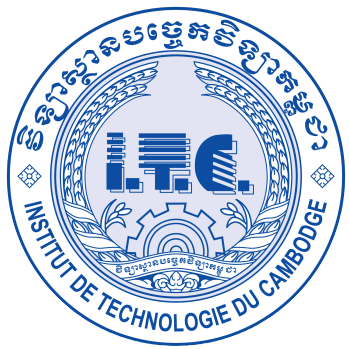
Regression Analysis

Chapter 03
Multiple Linear Regresson

Phauk Sokkhey
phauk.sokkhey@itc.edu.kh
Nhim Malai
nhim.malai@itc.edu.kh

Department of Applied Mathematics and Statistics
Institute of Technology of Cambodia



Introduction

- Multiple linear regression allows **more than one** regressors in a mean function (known as regression model).

Contents

Adding a regressor to a simple linear regression model

- A response Y and the simple regression model

$$E(Y|X_1 =_1) = \beta_0 + \beta_1 x_1$$

- Adding a second regressor into the model

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad (3.1)$$

- The main idea to add a second regressor X_2 is to explain the **remaining variation** of Y that has not been explained by X_1.

Example: United Nation Data

United Nation Data contains national health, welfare, and education statistics for 213 places. The data set consists of 7 variables:

- `region`: Region of the world.
- `group`: A factor with level `oecd` for countries that are members of OECD.
- `fertility`: Total fertility rate, the number of children per woman.
- `ppgdp`: Per capita gross domestic product in US dollars.
- `lifeExpF`: Female life expectancy, years.
- `pctUrban`: Percent urban.
- `infantMortality`: Infant deaths by age 1 year per 1000 live births.

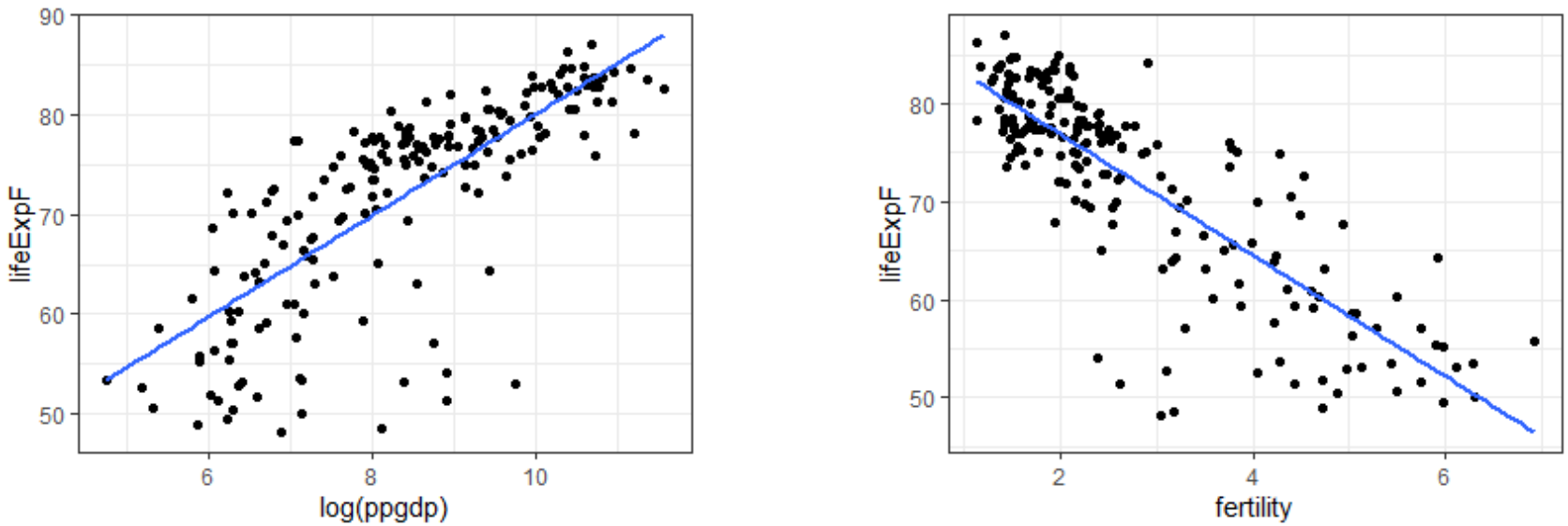


Figure 1: United Nations data on 199 locations, mostly nations: (left) `lifeExpF` versus `log(ppgdp)`; (right) `lifeExpF` versus `fertility`.

$$\hat{E}(\texttt{lifeExpF}|\texttt{log(ppgdp)}) = 29.26 + 5.09 \times \texttt{log(ppgdp)} \qquad (3.2)$$

- $R^2 = 0.5972$
- Ignores variable `fertility`

$$\hat{E}(\texttt{lifeExpF}|\texttt{fertility}) = 89.31 - 6.20 \times \texttt{fertility}$$

- $R^2 = 0.6787$
- Ignores variable `log(ppgdp)`

If the regressors `log(ppgdp)` and `fertility` were uncorrelated, then the marginal plots would provide a complete summary of the dependence of the response on the regressors, as the effect of `fertility` adjusted for `log(ppgdp)` would be the same as the effect of `fertility` ignoring `log(ppgdp)`.

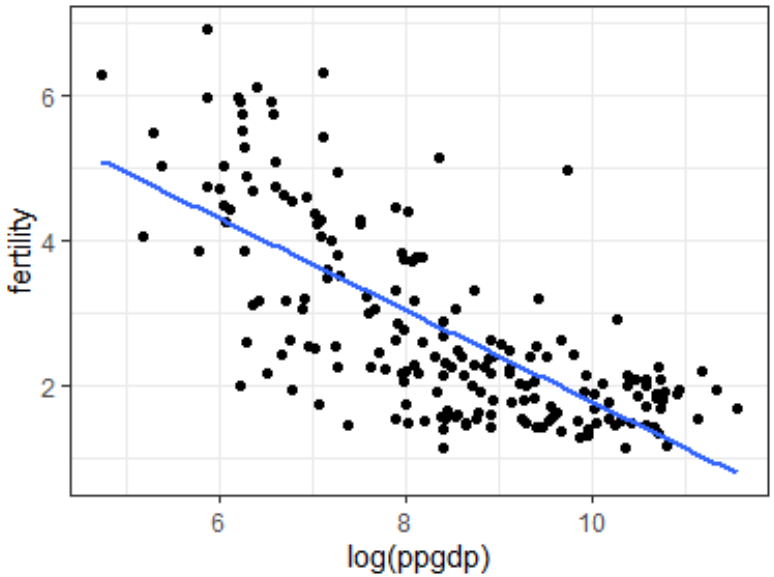


Figure 2: Marginal plot of `fertility` versus `log(ppgdp)`

.

- Countries with larger `log(ppgdp)` also tend to have lower `fertility`.
- `log(ppgdp)` and `fertility` are negatively correlated.
- These regressors partly explain the same variation.

Explaining Variability

What can be said about the proportion of variability in `lifeExpF` explain jointly by `log(ppgdp)` and `fertility`?

- The total explained variation (R^2) must be at least 67.87%.
- If the regressors were uncorrelated, then the variation explained by them jointly would be the sum of the variation explained individually ($59.92\% + 67.87\% = 127.78\% > 100\%$).
- From Figure 2, `log(ppgdp)` and `fertility` are correlated (correlation coefficient $r = -0.72$), thus the sum is greater than 100% and won't apply.
- The variation explained by both variables can be smaller than the sum of the individual variation explained if the regressors are partly explaining the same variation.

Added-Variable Plots

To get the effect of adding `fertility` to the model that already includes `log(ppgdp)`, we need to examine the univariate models for each regressor.

1. Compute the regression of the response `lifeExpF` on the first regressor `log(ppgdp)`. Keep the residuals from this regression. These residuals are part of the response `lifeExpF` *not explained* by `log(ppgdp)`.
2. Compute the regression of `fertility` on `log(ppgdp)`. Keep the residuals from this regression. These residuals are part of the new regressor `fertility` *not explained* by `log(ppgdp)`.
3. The *added-variable plot* is of the unexplained part of the response from (1) on the unexplained part of the added regressor from (2).

Example: United Nation Data

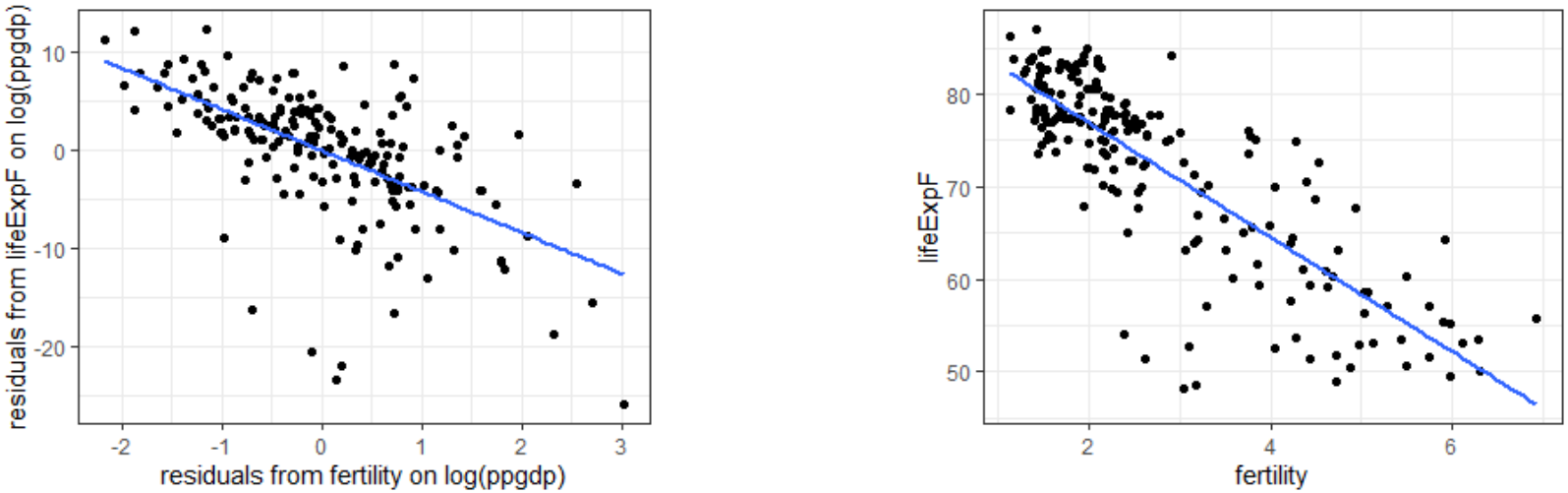


Figure 3: (left) Added-variable plot for `fertility` after `log(ppgdp)`. (right) The marginal plot of `lifeExpF` versus `fertility` ignoring `log(ppgdp)`.

- If Figure 3 (left) shows less variation about the fitted line than Figure 3 (right), then the two variables act jointly to explain extra variation.
- If the two graphs have similar variation, then the total explained variation by both variables is less than the additive amount.
- The latter is the case here.

R-code

```
lm1 = lm(lifeExpF~log_ppgdp, data=UN)
lm2 = lm(fertility~log_ppgdp, data=UN)
e1 = lm1$residuals
e2 = lm2$residuals
lm_e <- lm(e1~e2)
summary(lm_e)
```

Output

```
Call:
lm(formula = e1 ~ e2)

Residuals:
     Min       1Q   Median       3Q      Max
-22.6398  -1.7186   0.4222   2.6547  11.8321

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.557e-16  3.709e-01    0.00        1
e2          -4.177e+00  3.965e-01  -10.54   <2e-16

(Intercept)
e2          ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.153 on 191 degrees of freedom
Multiple R-squared:  0.3675,Adjusted R-squared:  0.3642
F-statistic:   111 on 1 and 191 DF,  p-value: < 2.2e-16
```

If we fit the simple linear regression to Figure 3 (left), we obtain:

$$\hat{E}(\hat{e}_1|\hat{e}_2) = 0 - 4.177 \times \hat{e}_2$$

where:

- $\hat{e}_1$: estimated residual from `lifeExpF` on `log(ppgdp)`.
- $\hat{e}_2$ estimated residual from `fertility` on `log(ppgdp)`.

The model has

- Estimated intercept $\hat{\beta}_0 = 0$.
- Estimated slope $\hat{\beta}_2 = -4.177$. This is exactly the estimated $\hat{\beta}_2$ that would be obtained when adding both regressors in a model.
- $R^2 = 0.367$ meaning that after adjusted for `log(ppgdp)`, adding `fertility` explains 36.7% of the remaining variability in `lifeExpF`.

Now we have two estimates of the coefficient β_2 for `fertility`:

- $\hat{\beta}_2 = -6.20$ ignoring `log(ppgdp)`
- $\hat{\beta}_2 = -4.177$ adjusted for `log(ppgdp)`

The slope in the added-variable plot is about 30% smaller than the slope in the plot that ignores `log(ppgdp)`, the effect of `fertility` is still important. The regressor `fertility` is useful after adjusting for `log(ppgdp)`.

Contents

The additive multiple linear regression model

The additive multiple linear regression model with response Y and regressors $X_1, ..., X_p$ will have the form:

$$E(Y|X) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \tag{3.3}$$

Or, when conditioning on specific values for the predictors $x_1, ..., x_p$ that we will collectively call $\boldsymbol{x}$:

$$E(Y|X = \boldsymbol{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{3.4}$$

Or equivalently,

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i \tag{3.3*}$$

with ε_i i.i.d. $N(0, \sigma^2)$.

Interpretation of the β-parameters

For parameter β_1 this follows from the following calculation:

$$E(Y|x_1+1, x_2, ..., x_p) - E(Y|x_1, x_2, ..., x_p)$$

$$= m(x_1+1, x_2, ..., x_p) - m(x_1, x_2, ..., x_p)$$
$$= (\beta_0 + \beta_1(x_1+1) + \beta_2 x_2 + ... + \beta_p x_p) - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)$$
$$= \beta_1$$

Hence, the parameter β_1 quantifies the increase in expected outcome when the regressor x_1 increases with one unit, while the other regressors in the model remain constant.

More generally, this also can be done for β_j.

Within the matrix notation, model 3.3* can be written as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (3.3^{**})$$

with $\boldsymbol{\varepsilon} \sim \boldsymbol{MVN}(\boldsymbol{0}, \boldsymbol{I_n}\boldsymbol{\sigma}^2)$ (MVN stands for multivariate normal distribution). With this notation we also can write $m(\boldsymbol{x})$ instead of $m(x_1, ..., x_p)$.

The parameters of model 3.3** can be estimated with the least squares estimation procedure. The method is identical to the method described for simple linear regression in Chapter 2.

All the results that were provided in Chapter 2 and that made use of the matrix notation, are still valid here. In the previous chapter the vector $\boldsymbol{\beta}$ had two elements (β_0 and β_1). Now the vector $\boldsymbol{\beta}$ has $p+1$ elements ($\beta_0, \beta_1, ..., \beta_p$). The design matrix had columns, whereas it now has $p+1$ columns.

All theories related to the sampling distributions, confidence and prediction intervals and hypothesis tests remain valid for the additive multiple linear regression model.

Predictors and regressors

Potential predictors could be:

- Continuous measurements: height, weight, ...
- Discrete and ordered measurements: doctor's rating of the overall health of a patient, ...
- Categorical variables: eye colors, ...
- Dummy variables and factors, ...

Some remedies to apply to variables to obtain better fit:

- Transformations of predictors: `log`-, `logit`-transformation, ...
- Polynomials: adding X^2, X^3, ... to the model
- Interactions: product of two or more variables
- Other combinations of predictors: BMI, sores, ...

Example: United Nation Data

For United Nation Data, we consider the model

$$\texttt{lifeExpF}_i = \beta_0 + \beta_1 \times \texttt{log(ppgdp)}_i + \beta_2 \times \texttt{fertility}_i + \varepsilon_i$$

with ε_i i.i.d. $N(0, \sigma^2)$.

R-code

```
Y <- UN$lifeExpF
XReg <- as.matrix(UN[,c("log_ppgdp", "fertility")])
head(XReg)
X<-cbind(1, XReg)
#head(X)
beta.hat <- solve(t(X)%*%X)%*%t(X)%*%Y
beta.hat
```

Output

```
                [,1]
          63.057988
log_ppgdp  2.452797
fertility -4.177172
```

The same estimates can be obtained with the `lm()` function in R:

R-code

```
lm_UN <- lm(LifeExpE ~ log_ppgdp + fertility, data=UN)
summary(lm_UN)
```

Output

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  63.0580     3.8216  16.500  < 2e-16***
log_ppgdp     2.4528     0.3481   7.046 3.29e-11***
fertility    -4.1772     0.3976 -10.507  < 2e-16***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.166 on 190 degrees of freedom
Multiple R-squared:  0.7452,Adjusted R-squared:  0.7426
F-statistic: 277.9 on 2 and 190 DF,  p-value: < 2.2e-16
```

Multiple linear regression summary

Table 1: Multiple linear regression summary in United Nation Data

	Estimate	**Std. Error**	***t*-Value**	**Pr(>|t|)**
(Intercept)	63.05	3.8216	16.500	<2e-16
log(ppgdp)	2.45	0.3481	7.046	3.29e-11
fertility	-4.17	0.3976	-10.507	<2e-16

$\hat{\sigma} = 5.1666$ with $190 df$, $R^2 = 0.7452$

Confident interval

R-Code

```
confint(lm_UN)
```

Output

```
                2.5 %    97.5 %
(Intercept) 55.519810 70.596166
log_ppgdp    1.766169  3.139425
fertility   -4.961367 -3.392978
```

Interpretation of the regression coefficients

Interpretation of regression coefficients:

- $\hat{\beta}_1 = 2.45$, we estimate that for a country with the same total fertility rate, female life expectancy increases on average by 2.45 years for an increase in the natural logarithm of per capita gross domestic product in the state. The effect of `log(ppgdp)` is the same for each fixed value of `fertility`.
- $\hat{\beta}_2 = -4.17$, we estimate that for a country with the same natural logarithm of per capita gross domestic product, female life expectancy drops on average by 4.17 years for an increase in total fertility rate (the number of children per woman). The effect of `fertility` is the same for each fixed value of `log(ppgdp)`.

Hypotheses Concerning On Coefficient

The multiple regression model has many regression coefficients, so many tests are possible. In this section, we consider only the testing of individual coefficients. The hypothesis tested is:

$$H_0 : \beta_j = 0 \qquad (j = 0, 1, ..., p')$$

$$H_a : \beta_j \neq 0 \qquad (j = 0, 1, ..., p')$$

As seen in simple regression, the distributions of the studentised LSE follow t-distribution with $n - p'$ degree of freedom where $p' = p + 1$ is the number of parameters in the model. Therefore, the hypothesis can be tested using this distribution.

Example: United Nation Data

For United Nation Data, we are interested in testing the hypothesis:

$$H_0 : \beta_j = 0 \qquad (j = 0, 1, 3)$$

$$H_a : \beta_j \neq 0 \qquad (j = 0, 1, 3)$$

The test statistics and p-values can be found in Table 1. All the p-values are significant at 5% level of significance. The table presents two-sided p-values.

A one-sided p-value can be obtained from a two-sided p-value for t-distribution if interested. For example, testing $\beta_1 < 0$, the one-sided p-value is obtained by 1 minus the two-sided p-value divided by 2 ($1 - 3.29\text{e-}11/2 \approx 1$) because $\hat{\beta}_1 > 0$. For the one-sided test that $\beta_1 > 0$, the one-sided p-value is obtained by the two-sided p-value divided by 2 ($3.29\text{e-}11/2 \approx 0$).

Prediction

Suppose we have a new case with its own set of predictors that result in a vector of regressor $\boldsymbol{x}$. We would like to predict the outcome given $\boldsymbol{x}$. We have proved the distribution of prediction error for simple linear regression in chapter 2 which is also valid here.

For example, for United Nation Data, we want to predict the female life expectancy a country with per capita gross domestic product of 50000USD (log(50000) = 10.82) and total fertility rate of 2.

R-code

```
predict(lm_UN, newdata = data.frame(log_ppgdp=10.81978,
                                    fertility=2),
        interval="prediction", level = 0.95)
```

Output

```
       fit      lwr     upr
1 81.24237 70.94503 91.5397
```

R-squared and Adjusted R-squared

- n: number of observations
- K: number of parameters (β_0, β_1,..., β_{K-1}) in the model
- RSS: Residual Sum of Squares, $\sum_{i=1}^{n}(y_i - \hat{y})^2$.
- TSS: Total Sum of Squares, $\sum_{i=1}^{n}(y_i - \bar{y})^2$.

R-squared

$$R^2 = 1 - \frac{RSS}{TSS}$$

Adjusted R-squared

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n-K)}{TSS/(n-1)}$$

- R^2 quantifies how well a model fits the data.
- R^2 always be improved when parameters are added to the model.
- Adjusted R^2 accounts for the number of parameters fit by the regression.
- Adjusted R^2 can be used to compare models with different numbers of parameters.

Contents

Interaction

We extend model our model by adding an **interaction term**. Consider the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{x_i 2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i \tag{3.5}$$

with ε_i i.i.d. $N(0, \sigma^2)$.

- $\beta_3 x_{i1} x_{i2}$ is the **interaction term** which quantifies the interaction of the regressor x_{i1} and x_{i2} on the mean outcome.
- We call the term $\beta_1 x_1$ and $\beta_2 x_{i2}$ the **main effects** of the regressor x_1 and x_2 respectively.

Interpretation of β parameters

We calculate the difference in expected outcome when x_1 increases with one unit when the regressor x_2 is kept constant:

$$E(Y|x_1+1, x_2) - E(Y|x_1, x_2)$$

$$= [\beta_0 + \beta_1(x_1+1) + \beta_2 x_2 + \beta_3(x_1+1)x_2] - [\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2]$$
$$= \beta_1 + \beta_3 x_2$$

This expression shows that the effect of x_1 depends on the value of x_2. This effect is the same for all values of x_2.

$$E(Y|x_1, x_2+1) - E(Y|x_1, x_2)$$

$$= [\beta_0 + \beta_1 x_1 + \beta_2(x_2+1) + \beta_3 x_1(x_2+1)] - [\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2]$$
$$= \beta_2 + \beta_3 x_1$$

Hypothesis testing

In model 3.5, we test the hypothesis in the following order:

1. test $H_0 : \beta_3 = 0$ (test for absence on interaction effect)
2. test $H_0 : \beta_1 = 0$ and/or $\beta_2 = 0$ (test for absence of main effects) only if there is no evidence or indication for the presence of an interaction effect.

Example: United Nation Data

R-code

```
lm_UN_Inter <- lm(lifeExpF ~ log_ppgdp*fertility, data=UN)
summary(lm_UN_Inter)
```

Output

```
Residuals:
     Min       1Q   Median       3Q      Max
-22.0802  -1.7589   0.4092   2.5904  11.9410

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         56.1356     5.4626  10.276  < 2e-16 ***
log_ppgdp            3.3782     0.6285   5.375 2.24e-07 ***
fertility           -1.3823     1.6328  -0.847   0.3983
log_ppgdp:fertility -0.3944     0.2236  -1.764   0.0793 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.138 on 189 degrees of freedom
Multiple R-squared:  0.7494,Adjusted R-squared:  0.7454
F-statistic: 188.4 on 3 and 189 DF,  p-value: < 2.2e-16
```

We obtained $\hat{\beta}_3 = -0.39$ meaning that the effect of `fertility`, decreases with an additional 0.39 years for each increase in `log(ppgdp)` (the effect of `fertility` is being $\beta_2 + \beta_3$`log(ppgdp)`).

The p-value for testing $H_0 : \beta_3 = 0$ versus $H_1 : \beta_3 \neq 0$ equals $p = 0.0793$. Thus, at the 5% level of significance, there is hardly evidence of an interaction effect, and if it were present the estimate and its confidence interval indicate only a very small effect.

Notation for Interaction Effect:
When the interaction effect is not significant, we adapt the convention to first remove the interaction term from the model and refit the model before looking at the main effects.

Interaction: Continuous vs. Dummy binary

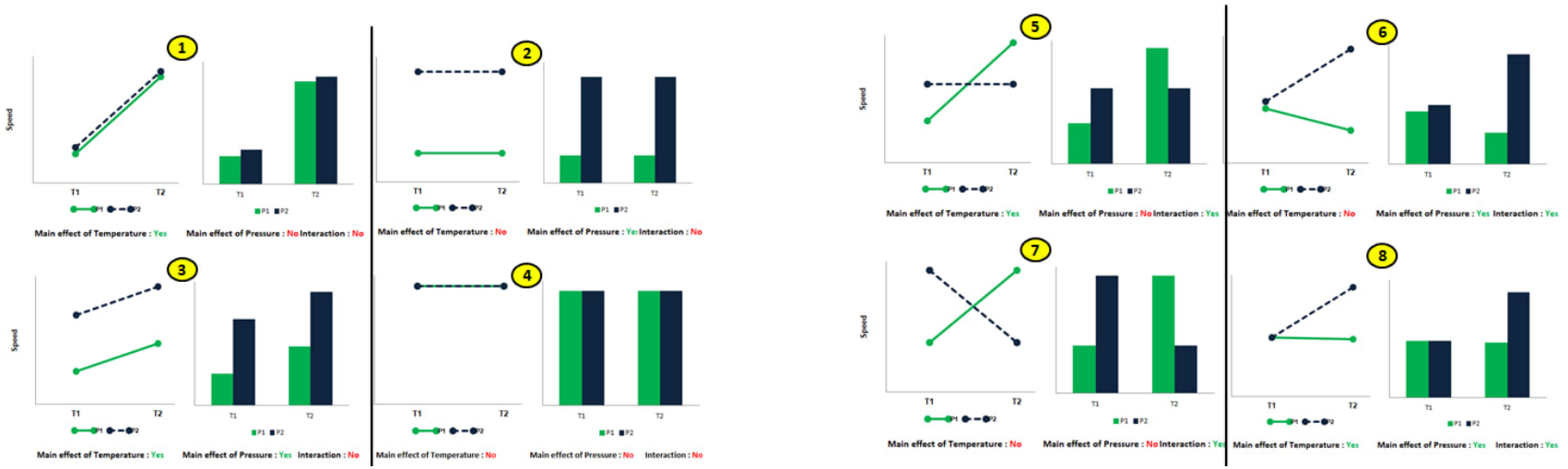


Figure 4: (Left) indicating **NO** interaction effect and (right) indicating **THERE IS** interaction effect

Source: Main Effects Plot

Example: United Nation Data

For an illustration, let's make a new binary variable from `fertility` where `bin.fert` is 1 when `fertility` is more than 3 and otherwise 0.

R-Code

```
library(dplyr)
UN = mutate(UN, bin_fert = if_else(fertility >3, 1, 0))
library(ggplot2)
UN$bin_fert = as.factor(UN$bin_fert)
ggplot(UN, aes(x=log_ppgdp, y=lifeExpF)) +
  geom_point(aes(color=bin_fert)) +
  geom_smooth(aes(group=bin_fert),
              method = "nls", formula = y ~ a * x + b, se = FALSE,
              method.args = list(start = list(a = 0.1, b = 0.1))) +
  theme_bw()
```

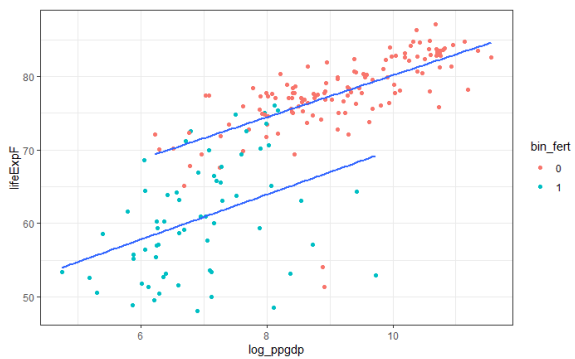


Figure 5: Scatter plots and regression lines of `log(ppgdp)` by `bin.fert`

Likelihood Ratio Test for Interaction Effect

- Full model (Model with an interaction effect):

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{1i} x_{2i} + \varepsilon_i$$

- Nested mode (Model without an interaction effect):

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

- Hypothesis:
  - H_0: The two models are the same (No interaction effect)
  - H_A: The two models are different (There is an interaction effect)

$$\text{LRT} = -2 \log \left(\frac{L_{\text{nested model}}(\hat{\theta})}{L_{\text{full model}}(\hat{\theta})} \right) \sim \chi^2_{df=(df \text{ from nested model}) - (df \text{ from full model})}$$

Example: United Nation Data

R-Code

```
#likelihood ratio test for interaction
LRT.int = 2*(logLik(lm_UN_Inter) - logLik(lm_UN))
c(LRT.int, 1-pchisq(LRT.int,1))

#alternative way
library(lmtest)
lrtest(lm_UN_Inter, lm_UN)
```

Output

```
[1] 3.15233557 0.07581837


Likelihood ratio test

Model 1: lifeExpF ~ log_ppgdp * fertility
Model 2: lifeExpF ~ log_ppgdp + fertility
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   5 -587.70
2   4 -589.28 -1 3.1523    0.07582 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Contents

Multicollinearity

- Suppose $\boldsymbol{X}$ *isthedatamatrixforthesetofregressors*
- We say that the set of regressors is *collinear* is we can find a vector of constants $\boldsymbol{a}$ such that $\boldsymbol{Xa} \approx \boldsymbol{0}$.
- If the "$\approx$" is replaced by an "$=$" sign, then at least one of regressors is a linear combination of the others and we have overparameterized model.
- If $\boldsymbol{X}$ is collinear, then the R^2 for the regression of one of the regressors on all the remaining regressors, including the intercept, is close to one.
- (Multi)collinearity depends on the sample correlations between the regressors, not on theoretical population quantities.

Example: Water Usage in Minnesota

The data file `MinnWater` provides yearly water consumption in Minnesota from 1988-2011.

- `year`: year
- `allUse`: total groundwater consumption, statewide, in billions of gallons
- `muniUse`: total municipal water consumption, statewide, in billions of gallons
- `irrUse`: consumption for irrigation in 13 counties, in billions of gallons `agPrecip`: average growing season June to August precipiciation (inches) for the 13 Minnesota counties that use the most irrigation `muniPrecip`: average May to September precipiciation (inches) for the 10 Minnesota counties with highest municipal water pumping
- `statePop`: estimated state population
- `muniPop`: estimated 10 county urban population

For water usage in Minnesota example, we consider the response variable is `log(muniUse)`, and potential predictors are `yeat` and `muniPrecip` and `log(muniPop)`.

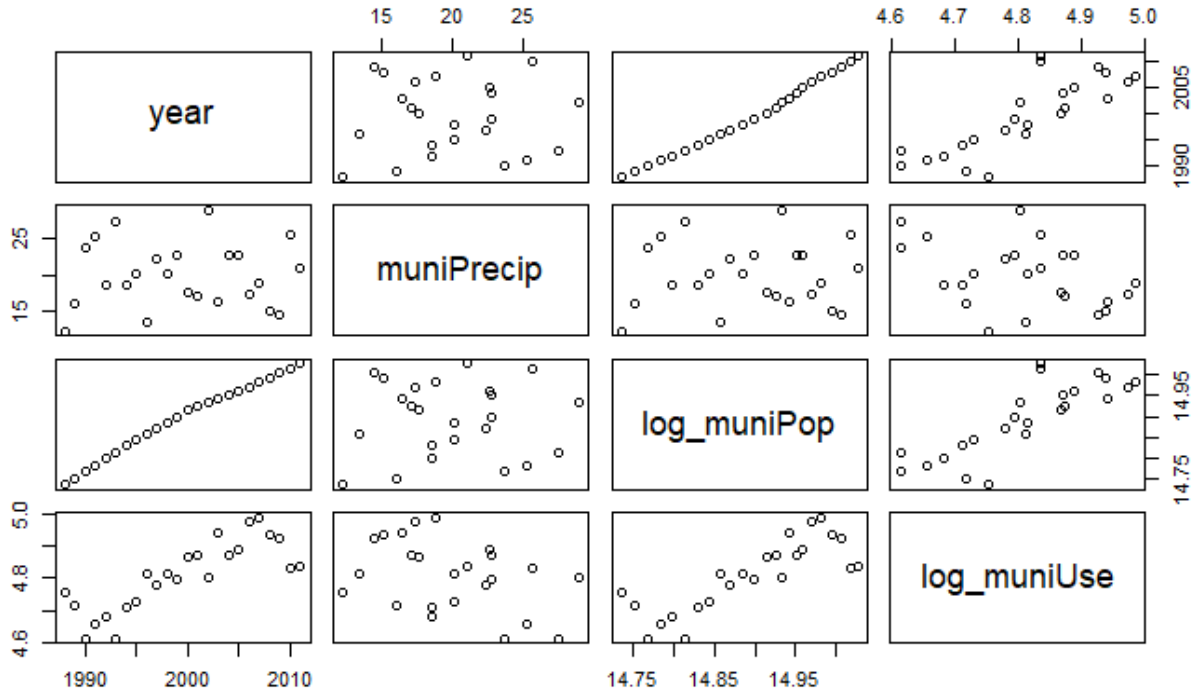


Figure 6: Scatterpot matrix for the Minnesota water use data

Table 2: Regression of `log(muniUse)` on Different Combinations of Regressors for the Minnesota Water Use Data

Regressor	Model 1	Model 2	Model 3
(Intercept)	−20.0480*	−20.1584*	−1.2784
`year`	0.0124*	0.0126*	−0.0111
`muniPrecip`		−0.0099*	−0.0106*
`log(muniPop)`			1.9174

*Indicates p-value < 0.01.

- **Model 1**: As expected, `log(muniUse)` is increasing overtime.
- **Model 2**: When `muniPrecip` is added, the estimate of for `year` hardly changes, *as expected from the lack of correlation between* `year` *and* and `muniPrecip`.
- **Model 3**: Adding `log(muniPop)`, however, tells a different story: the coefficient for `year` is much smaller and negative, and is no longer significant. The cause of this is clear: `log(muniPop)` is highly correlated with `year`. These two variables explain the sample variation in `log(muniUse)`.

Variance Inflation Factor (VIF)

The **Variance Inflation Factor (VIF)** can be used for measuring the effect of multicollinearity on the variances (or standard errors) of the parameter estimators.

- $R_j^2 = 0$: the variability of regressor j cannot be explained by a linear model with other regressors acting as regressors (regressor j is linearly independent of the other regressors).
- $R_j^2 = 1$: the variability of regressor j can be explained by a linear model of other regressors (regressor j is a linear combination of the other regressors for all n sample observations)

Variance Inflation Factor (VIF)

The variance of the estimated coefficient of the multiple linear regression model is defined by:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} \times \frac{1}{1 - R_j^2}$$

where σ^2 is the residual variance of the model that includes all the $(p-1)$ regressors.

From the equation above, $Var(\hat{\beta}_j)$ is the product of:

1. the variance of $\hat{\beta}_j$ in a model without collinearity ($R_i^2 = 0$) for regressor j
2. the variance inflation factor (VIF)

When there is a problem and what to do:

- Some guidelines give VIF $= 10$ as a threshold and others give VIF $= 5$ for a large problematic value for VIF.
- When there is a large problematic VIF, you may want to remove one or more regressors from the model (if this is allowed and makes sense keeping the research question in mind).

Example: Water Usage in Minnesota

R-code

```
R2.year <- summary(lm(year~muniPrecip+log_muniPop, data=waterMinn))$r.squared
R2.muniPrecip <- summary(lm(muniPrecip~year+log_muniPop, data=waterMinn))$r.squared
R2.log_muniPop <- summary(lm(log_muniPop~ year + muniPrecip, data=waterMinn))$r.squared
1/(1-R2.year)
1/(1-R2.muniPrecip)
1/(1-R2.log_muniPop)
```

Output

```
[1] 116.9698
[1] 1.032013
[1] 117.0957
```

R-code

```
vif(lm(log_muniUse ~ year + muniPrecip + log_muniPop, data = waterMinn))
```

Output

```
       year  muniPrecip log_muniPop
 116.969782    1.032013  117.095745
```

Example: Water Usage in Minnesota

Output

```
      year  muniPrecip log_muniPop
116.969782    1.032013  117.095745
```

- The VIF of `muniPrecip` is very close to 1, and hence is not problematic at all.
- the VIF of `year` and `log(muniPop)` is very large (greater than 5 or 10), hence is problematic. Consider removing one of the variables.
- The VIF ≈ 117 meaning that the variance of the parameter estimator is 117 times larger than if there were no multicollinearity and hence the standard error is $\sqrt{117} = 10.81$ larger as compared to no multicollinearity.

Contents

Leverage

Leverage as a tool to measure the influence of an observation on the regression fit.

- It is not advised for the statistician to remove observations only because they were identified as outliers.
- It is the task of the statistician to identify outliers and to report them so that the scientists who are closer to the data and the study can check whether perhaps something went wrong that may explain the outlying behaviour of the data point.
- Sometimes, an outlier does not strongly affect the parameter estimates and the conclusions, thus such outliers are usually not problematic.
- Other times an outlier may be very **influential** in the sense that this outlying observation has a strong effect on the numerical values of the parameter estimates, thus such outliers are problematic and worrisome.

Leverage

In matrix notation, the vector of predictions can be written as

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y},$$

where the $n \times n$ matrix $\boldsymbol{H}$ is generally known as the **hat-matrix**. Note that the hat-matrix is *idempotent*, i.e.

$$\boldsymbol{H}\boldsymbol{H}^t = \boldsymbol{H}^t\boldsymbol{H} = \boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}.$$

The ith element of $\hat{\boldsymbol{Y}}$, i.e. $\hat{Y}_i$, can be written as

$$\sum_{j=1}^{n} h_{ij} Y_j$$

with h_{ij} the element on position (i, j) of matrix $\boldsymbol{H}$. This equation demonstrates that the predictions are linear functions of the outcomes (it is an example of a linear predictor).

Without proof we give here the following property:

$$\sum_{j=1}^{n} h_{ij} = 1 \text{ for all } i = 1, ..., n$$

For the prediction of observation i we can write

$$\hat{Y}_i = \boldsymbol{h}_i^t \boldsymbol{Y}$$

with $\boldsymbol{h}_i^t$ the ith row of H. Since the sum of the elements of $\boldsymbol{h}_i$ always equals 1, the above equation show that the prediction $\hat{Y}_i$ is a weighted mean of the sample outcomes $Y_1, ..., Y_n$.

This interpretation allows us to evaluate the elements of the vector $\boldsymbol{h}_i$:

- if h_{ij} is large (relative to the other elements), then outcome Y_j strongly affects the prediction $\hat{Y}_i$.
- A global measure for the influence of observation Y_i on the predictions $\hat{Y}_1, ..., \hat{Y}_n$ is given by

$$\sum_{j=1}^{n} h_{ij}^2 = \boldsymbol{h}_j^t \boldsymbol{h}_i = h_{ii}$$

(the final equality follows from $\boldsymbol{HH} = \boldsymbol{H}$). The square ($h_{ij}^2$) is used because both large positive and large negative h_{ij} imply that Y_i is influential.

Leverage

- The **leverage** of observation i is defined as h_{ii} and it is thus a global measure of the influence of the observation i on the predictions.
- It can also show that $\sum_{i=1}^{n} h_{ii} = p$ where p is the number of parameter in a model. This may help in thresholding the individual h_{ii} leverage value i.e. the average of the h_{ii} is thus given by p/n. Leverages much larger than p/n may be called influential.
- If an observation i is identified as an outlier, as if its leverage h_{ii} is large, then we call observation i an **influential outlier**.

Example: United Nation Data

R-code

```
m<-lm(lifeExpF ~ fertility + log_ppgdp, data=UN, x=T)
X<-m$x
H<-X%*%solve(t(X)%*%X)%*%t(X)
h<-diag(H)

#OR simply way of computing the leverages:
h<-influence(m)$h

sum(h)

plot(h, xlab="i", ylab="leverage", cex.axis=1.5, cex.lab=1.5)
abline(h=sum(h)/nrow(UN), lty=2)
```

Example: United Nation Data

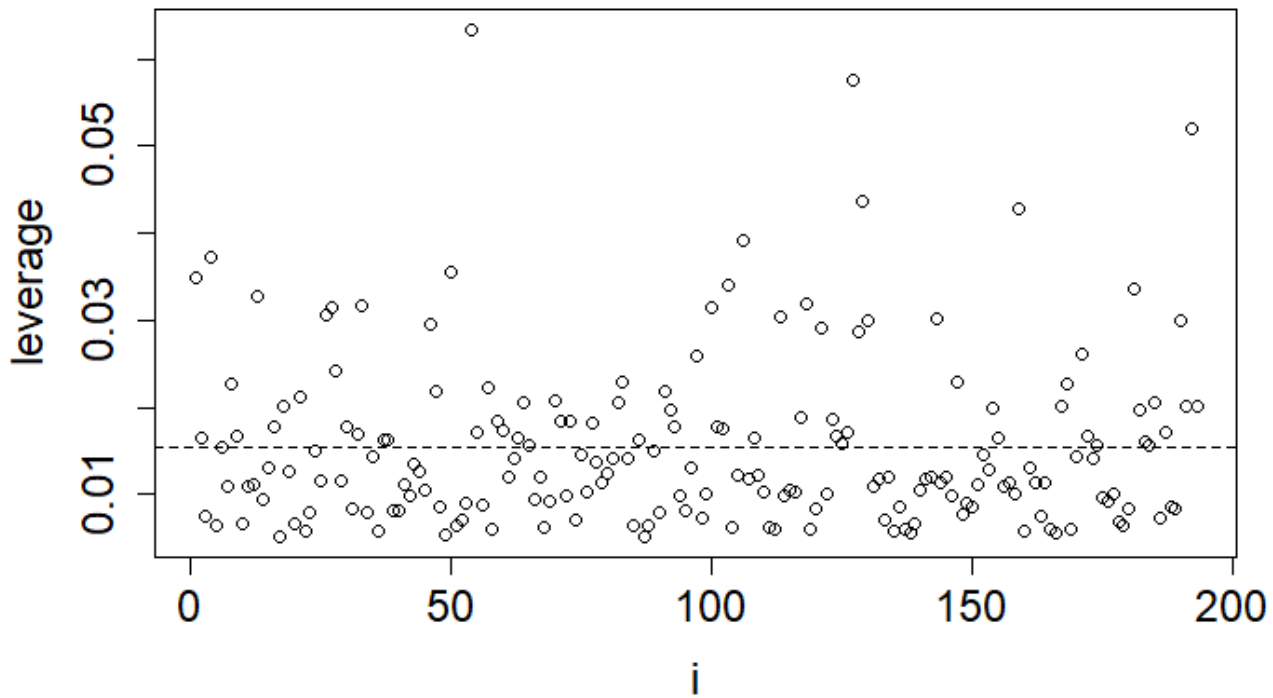


Figure 7: Leverage plot of each observation

The leverage plot does not indicate any strong influential outliers.

Contents