

## 13 Multiple Regression

MTH 3240 Environmental Statistics

Spring 2020

### Objectives

Objectives:

- State and interpret the multiple regression model (**Optional for Spring 2020**).
- Obtain and interpret estimates of multiple regression model coefficients.
- Obtain and interpret fitted values and residuals associated with a fitted multiple regression model.
- Interpret the  $R^2$  associated with a fitted multiple regression model.
- Carry out  $t$  tests for the coefficients in a multiple regression model.

### Introduction to Multiple Regression

- Often a **single** explanatory variable doesn't adequately explain the variation in the response variable.  
Instead, **multiple** explanatory variables are needed.
- **Multiple regression analysis** refers to fitting a regression model containing **multiple** explanatory variables.

- Some reasons for including **more than one** explanatory variable in a model:
  - The model may explain substantially more of the variation in  $Y$  than one containing a single explanatory variable, thereby giving **better predictions** of  $Y$  values.
  - The model allows us to study the effect of one explanatory variable on  $Y$  while **controlling** for the effects of the others.

Notes

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

## Example

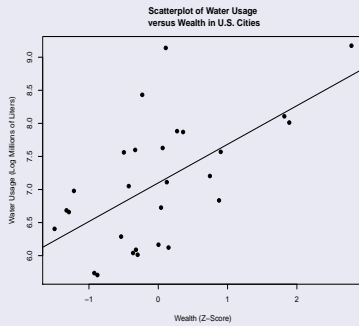
The data below show, for  $n = 28$  U.S. cities, the **water consumption** (log of millions of liters/day), **population** (in millions in 2000), and **wealth** ( $z$ -score of the city's median income).

Water Usage for U.S. Metropolitan Areas			
City	Water Usage ( $Y$ )	Wealth ( $X_1$ )	Population ( $X_2$ )
New York	9.17	2.787	21.286
Los Angeles	9.14	0.108	16.374
Chicago	8.43	-0.231	9.158
DC/Baltimore	8.11	1.819	6.484
San Francisco	8.01	1.890	6.263
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Stockton	5.74	-0.923	0.564
Mobile	6.41	-1.496	0.540

MTH 3240 Environmental Statistics

Multiple Regression

Here's a scatterplot of the **water consumption** ( $Y$ ) vs **wealth** ( $X$ ).



MTH 3240 Environmental Statistics

Multiple Regression

The equation of the **fitted regression line** in the scatterplot is

$$\hat{Y} = 7.10 + 0.58X.$$

The positive slope indicates that **water consumption** increases by **0.58** units for each one-unit increase in **wealth**.

The **predicted water consumption** for a city whose **wealth** is, say, **1.0** is

$$\hat{Y} = 7.10 + 0.58(1.0) = 7.68.$$

MTH 3240 Environmental Statistics

Multiple Regression

Two reasons why **population** should be included as an explanatory variable in the model along with **wealth**:

- First, a city's **water consumption** is related to its **population size**, and differences among **population sizes** inflate the residuals in the scatterplot (two slides back).

In other words, a city's **wealth alone** doesn't **predict** its **water consumption** very well.

Including **population** in the model will lead to smaller residuals and better **predictions** of **water consumption**.

MTH 3240 Environmental Statistics

Notes

---



---



---



---



---



---



---



---

Notes

---



---



---



---



---



---



---



---

Notes

---



---



---



---



---



---



---



---

Notes

---



---



---



---



---



---



---



---

- Second, including **population** in the model will allow us to **control** for the effect of **population size** on **water consumption** while investigating the effect of **wealth**.

This is important because it turns out that **wealthier** cities tend to also be **larger**, so the effects of **wealth** and **population size** are **confounded**.

We'll see that including **population** in the model will allow us to investigate the effect of **wealth** on **water consumption** while holding **population size constant** (i.e. **controlling** for it).

Notes

---

---

---

---

---

---

---

---

### Example (Cont'd)

The so-called **fitted multiple regression model**, with **water consumption** as the response variable ( $Y$ ) and **both wealth** ( $X_1$ ) and **population size** ( $X_2$ ) as explanatory variables, is

$$\hat{Y} = 6.48 + 0.11 X_1 + 0.16 X_2$$

(obtained using software).

This is the **equation** of a **plane** in a three-dimensional coordinate system, as shown on the next slide.

Notes

---

---

---

---

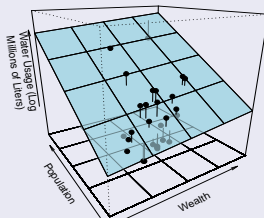
---

---

---

---

3D Scatterplot of Water Usage  
with Regression Surface



Notes

---

---

---

---

---

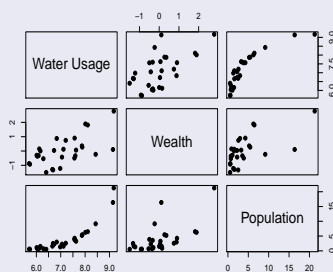
---

---

---

Here's a **scatterplot matrix** of the data.

Scatterplot Matrix of Water Usage,  
Wealth, and Size of a City



Notes

---

---

---

---

---

---

---

---

- In general, an equation of the form

$$Y = b_0 + b_1 X_1 + b_2 X_2$$

describes a **plane** relating a response variable  $Y$  to *two* explanatory variables  $X_1$  and  $X_2$ :

- The **intercept**  $b_0$  is the value of  $Y$  when  $X_1$  and  $X_2$  are **both zero**.
- The **coefficient**  $b_1$  quantifies the **change** in  $Y$  for each **one-unit change** in  $X_1$ , **while  $X_2$  is held constant**.
- The **coefficient**  $b_2$  quantifies the **change** in  $Y$  for each **one-unit change** in  $X_2$ , **while  $X_1$  is held constant**.

Notes

---

---

---

---

---

---

---

---

- The equation can be used to **predict** the value of  $Y$  from given values of  $X_1$  and  $X_2$  (by plugging the  $X_1$  and  $X_2$  values into the equation).

Notes

---

---

---

---

---

---

---

---

#### Example (Cont'd)

Here's the equation of the **fitted multiple regression model** again,

$$\hat{Y} = 6.48 + 0.11 X_1 + 0.16 X_2,$$

where  $Y$  is **water consumption**,  $X_1$  is **wealth**, and  $X_2$  is **population size**.

The **estimated coefficient** for **wealth** is  $b_1 = 0.11$ .

This says that **water consumption** increases by **0.11** units for each one-unit increase in **wealth**, **holding population size constant** (i.e. **controlling** for it).

In other words, the difference in **water consumptions** of two **same-sized** cities whose **wealths** differ by **one unit** would be about **0.11** units.

Notes

---

---

---

---

---

---

---

---

By contrast, in the model that **only** contained **wealth**, the slope **0.58** indicated that when we **don't control for population size**, the **water consumption** increases by **0.58** units for each **one-unit** increase in **wealth** (more than five times as much as when we control for population size!)

Notes

---

---

---

---

---

---

---

---

The **predicted water consumption** for a city whose **wealth** score is **1.0** and whose **population** is **3.0 million** is

$$\hat{Y} = 6.48 + 0.11(1.0) + 0.16(3.0) = 7.07.$$

- When the response variable  $Y$  is related to **multiple** explanatory variables  $X_1, X_2, \dots, X_p$ , an equation of the form

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p$$

relates  $Y$  to those explanatory variables.

- The **intercept**  $b_0$  is the value of  $Y$  when  $X_1, X_2, \dots, X_p$  are **all zero**.
- Each **coefficient**  $b_k$  (for  $k = 1, 2, \dots, p$ ) quantifies the **change** in  $Y$  for a **one-unit change** in  $X_k$ , **while the other  $X$ 's are all held constant**.

- The equation can be used to **predict** the value of  $Y$  from given values of  $X_1, X_2, \dots, X_p$  (by plugging the  $X_1, X_2, \dots, X_p$  values into the equation).

### Example

Efficient design of municipal waste incinerators requires knowing the energy content of the waste.

In a study of the relationship between the **energy content** of waste and its **composition**, the following variables were measured on each of  $n = 30$  waste specimens:

$$Y = \text{Energy content (kcal/kg)}$$

$$X_1 = \text{Percent plastics by weight}$$

$$X_2 = \text{Percent paper by weight}$$

$$X_3 = \text{Percent garbage by weight}$$

$$X_4 = \text{Percent moisture by weight}$$

Notes

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

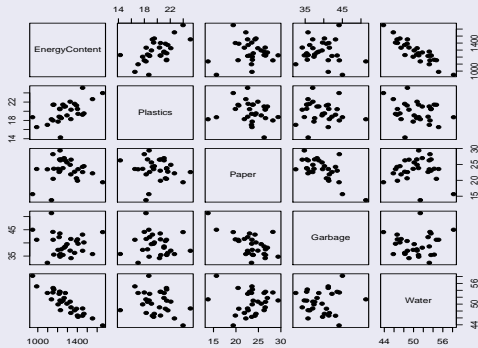
---

---

The data are below.

Waste Specimen	Municipal Waste Composition				
	Energy Content	Plastics	Paper	Garbage	Water
1	947	18.69	15.65	45.01	58.21
2	1407	19.43	23.51	39.69	46.31
3	1452	19.24	24.23	43.16	46.63
4	1553	22.64	22.20	35.76	45.85
5	989	16.54	23.56	41.20	55.14
6	1162	21.44	23.65	35.56	54.24
⋮	⋮	⋮	⋮	⋮	⋮
29	1391	21.25	20.63	40.72	48.67
30	1372	21.62	22.71	36.22	48.19

Scatterplot Matrix of Municipal Waste Data



The **fitted multiple regression model** (obtained using software) is

$$\hat{Y} = 2245 + 28.9 X_1 + 7.64 X_2 + 4.30 X_3 - 37.4 X_4.$$

The **estimated coefficient** for **plastics** ( $X_1$ ) is  $b_1 = 29.8$ .

This says the **energy content** increases by **29.8** units for each one-unit increase in **plastics**, *holding the other explanatory variables constant* (i.e. *controlling* for those variables).

The **estimated coefficient** for **paper** ( $X_2$ ) is  $b_2 = 7.64$ .

This says the **energy content** increases by **7.64** units for each one-unit increase in **paper** (*holding the other explanatory variables constant*).

The **estimated coefficient** for **garbage** ( $X_3$ ) is  $b_3 = 4.30$ .

This says the **energy content** increases by **4.30** units for each one-unit increase in **garbage** (*holding the other explanatory variables constant*).

The **estimated coefficient** for **water** ( $X_4$ ) is  $b_4 = -37.4$ .

This says the **energy content** *decreases* by **37.4** units for each one-unit increase in **water** (*holding the other explanatory variables constant*).

Notes

---

---

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

---

---

- The **response variable** is denoted by  $Y$  and the **explanatory variables** by  $X_1, X_2, \dots, X_p$ .
- We store the data in columns as below.

Observation	$Y$	$X_1$	$X_2$	$\dots$	$X_p$
1	$Y_1$	$X_{11}$	$X_{21}$	$\dots$	$X_{p1}$
2	$Y_2$	$X_{12}$	$X_{22}$	$\dots$	$X_{p2}$
3	$Y_3$	$X_{13}$	$X_{23}$	$\dots$	$X_{p3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n$	$Y_n$	$X_{1n}$	$X_{2n}$	$\dots$	$X_{pn}$

- Thus

$Y_i$  = The value of the response variable for the  $i$ th individual.

$p$  = The number of explanatory (predictor) variables.

$n$  = The number of individuals upon which the response and explanatory variables are measured, i.e. the sample size.

$X_{1i}, X_{2i}, \dots, X_{pi}$  = The values of the  $p$  explanatory variables for the  $i$ th individual.

## The Multiple Regression Model (Optional for Spring 2020)

- We'll describe the relationship between a **response variable** and **multiple explanatory variables** using a (theoretical) **statistical model** called the **multiple regression model**.

Later, we'll test **hypothesis** about the model **coefficients**.

## (Optional for Spring 2020)

- The model (next slide) reflects a true (unknown) **non-random relationship** between the response and explanatory variables, but allows for **random deviations** away from that relationship.

## Notes

---

---

---

---

---

---

---

---

---

---

## Notes

---

---

---

---

---

---

---

---

---

---

## Notes

---

---

---

---

---

---

---

---

---

---

## Notes

---

---

---

---

---

---

---

---

---

---

(Optional for Spring 2020)

**Multiple Linear Regression Model with  $p$  Explanatory Variables:**

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi} + \epsilon_i,$$

where

$Y_i$  is the observed value of the response variable for the  $i$ th individual ( $i = 1, 2, \dots, n$ ).

(Optional for Spring 2020)

$X_{1i}, X_{2i}, \dots, X_{pi}$  are the observed values of the  $p$  explanatory variables for the  $i$ th individual.

$\beta_0$  is the true (unknown) ***y-intercept*** of the underlying ***true regression model***.

$\beta_1, \beta_2, \dots, \beta_p$  are the true (unknown) ***coefficients*** for  $X_1, X_2, \dots, X_p$  in the ***true regression model***.

$\epsilon_i$  is a random error term following a  $N(0, \sigma)$  distribution (and the  $\epsilon_i$ 's are independent of each other).

(Optional for Spring 2020)

- When there are only **two** explanatory variables ( $p = 2$ ), the model is:

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

which describes a **plane**, but allows each  $Y_i$  to **deviate** above or below it by a **random** amount  $\epsilon_i$ .

Fitted Values and Residuals

- We **fit** the **model** to the data using the **method of least squares** (via statistical software).
- The **fitted multiple regression model** will be denoted by

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p.$$

The ***y-intercept***  $b_0$  and ***coefficients***  $b_1, b_2, \dots, b_p$  (obtained using statistical software) are called the ***least squares estimates*** of the true (unknown) ***population (or model) coefficients***, which are denoted  $\beta_0$  and  $\beta_1, \beta_2, \dots, \beta_p$ .

Notes

---



---



---



---



---



---



---

Notes

---



---



---



---



---



---



---

Notes

---



---



---



---



---



---



---

Notes

---



---



---



---



---



---



---



- (For the water consumption data,  $\beta_0$  and  $\beta_1$  and  $\beta_2$  would be the  $y$ -intercept and coefficients of the plane relating **water consumption** to **wealth** and **population size** in the **population** of cities.)

- The **fitted values**, denoted  $\hat{Y}_i$ , are values of the fitted model corresponding to **observed** values of **observed** values of  $X_1, X_2, \dots, X_p$ .

**Fitted Value:** For the  $i$ th individual in the data set,

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i} + \dots + b_pX_{pi},$$

where  $X_{1i}, X_{2i}, \dots, X_{pi}$  are the values of the  $p$  explanatory variables for that individual.

When there are only **two** explanatory variables ( $p = 2$ ), the **fitted values** lie on the **fitted plane**.

The **fitted values** are the  $Y$  values we'd **predict** for individuals in the **data set** that the model was fitted to.

- A **residual**, denoted  $e_i$ , is the difference between an **observed**  $Y_i$  value and that individual's **fitted value**  $\hat{Y}_i$ .

**Residual:** For the  $i$ th individual in the data set,

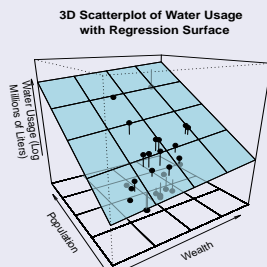
$$e_i = Y_i - \hat{Y}_i,$$

where  $Y_i$  is the observed response for that individual and  $\hat{Y}_i$  is the fitted value.

When there are only **two** explanatory variables ( $p = 2$ ), the **residuals** are the **deviations** above or below the fitted plane.

### Example

For the study of **water consumption**, **wealth**, and **population size** in  $n = 28$  cities, the **residuals** are the vertical **deviations** away from the **fitted plane**, and the **fitted values** are the contact points where the deviation lines meet the plane.



### Notes

---

---

---

---

---

---

---

---

### Notes

---

---

---

---

---

---

---

---

### Notes

---

---

---

---

---

---

---

---

### Notes

---

---

---

---

---

---

---

---

- A **fitted multiple regression model** represents the **effect of  $X_1, X_2, \dots, X_p$  on  $Y$** .

The **residuals** represent the **net effect of all other variables besides  $X_1, X_2, \dots, X_p$  on  $Y$** .

- **Example:** In the water consumption regression analysis, a **residual** represents the net effects of **all other variables besides wealth ( $X_1$ ) and population size ( $X_2$ ) on a city's water consumption ( $Y$ )** (e.g. its temperature, precipitation, landscape characteristics, number of swimming pools, etc.).

## R-Squared

- The **Multiple R-squared** (more formally, **coefficient of multiple determination**), denoted  $R^2$ , is a statistic that measures **how well the fitted model fits the data**.

(We'll see how  $R^2$  is computed later.)

- $R^2$  is interpreted as the **proportion of variation in  $Y$  that is explained by  $X_1, X_2, \dots, X_p$** :
  - An  $R^2$  value **close to one** means most of the  $Y$  variation is explained by the variables  $X_1, X_2, \dots, X_p$ , and the **model fits the data well**.

This shows up as **small residuals**.

- An  $R^2$  value **close to zero** means very little or none of the  $Y$  variation is explained by the variables  $X_1, X_2, \dots, X_p$  (but is explained by *other variables besides  $X_1, X_2, \dots, X_p$* ), and the **model doesn't fit the data**.

This shows up as **large residuals**.

### Example

For the study of **water consumption, wealth, and population size** of  $n = 28$  U.S. cities, when *both wealth and population* are included in the model,

$$R^2 = 0.747$$

(obtained using software).

Thus **74.7%** of the variation in cities' **water consumptions** is attributable to differences among their **wealths and population sizes**.

The other **25.3%** is due to the combined effects of **all other variables** (e.g. temperature, precipitation, landscape characteristics, number of swimming pools, etc.).

## Notes

---

---

---

---

---

---

---

---

---

---

---

---

## Notes

---

---

---

---

---

---

---

---

---

---

---

---

## Notes

---

---

---

---

---

---

---

---

---

---

---

---

## Notes

---

---

---

---

---

---

---

---

---

---

---

---

- $R^2$  is computed using **sums of squares**:

**Multiple R-Squared** (or **Coefficient of Multiple Determination**):

$$R^2 = \frac{SSR}{SSTo} \quad \left( = 1 - \frac{SSE}{SSTo} \right).$$

(We'll see later how the sums of squares SSR, SSTo, and SSE are computed.)

- SSR measures variation in  $Y$  due to  $X_1, X_2, \dots, X_p$ , and SSTo measures *total* variation in  $Y$ , so  $R^2$  is

$$R^2 = \frac{\text{Variation in } Y \text{ Due to } X_1, X_2, \dots, X_p}{\text{Total Variation in } Y}.$$

Thus  $R^2$  is interpreted as **the proportion of variation in  $Y$  that's explained by  $X_1, X_2, \dots, X_p$ .**

- In the **fitted regression model**

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p,$$

for each  $k = 1, 2, \dots, p$ , the coefficient  $b_k$  represents the (estimated) **change** in  $Y$  for a **one-unit increase** in  $X_k$  (*holding the other  $X$  variables constant, i.e. controlling for them*).

- If a  $b_k$  was **zero**, there'd be **no change** in  $Y$  for any given change in  $X_k$ , i.e. **no relationship** between  $Y$  and  $X_k$  (*controlling for the other  $X$  variables in the model*).

A  $b_k$  **different from zero** would mean there's a **relationship** between  $Y$  and  $X_k$ .

- But a  $b_k$  can differ from zero due to **sampling error** (because it's just an **estimate** based on data **sampled** from the population).
- For each explanatory variable  $X_1, X_2, \dots, X_p$ , we'll test the **null hypothesis** that there's **no relationship** between  $X_k$  and  $Y$ .

$$H_0 : \beta_k = 0$$

where  $\beta_k$  is the true (unknown) **population** (or **model**) coefficient for  $X_k$ .

---

---

---

---

---

---

---

---

---

---



---

---

---

---

---

---

---

---

---

---



---

---

---

---

---

---

---

---

---

---



---

---

---

---

---

---

---

---

---

---

- The alternative is that there's a **relationship** between  $X_k$  and  $Y$ .

$$H_a : \beta_k \neq 0$$

A separate ***t* test** is carried out for each of the explanatory variables  $X_1, X_2, \dots, X_p$ .

***t* Test Statistic for a coefficient:**

$$t = \frac{b_k - 0}{S_{b_k}},$$

where  $S_{b_k}$  is the (estimated) **standard error** of  $b_k$  (computed using statistical software).

- $t$  indicates how many **standard errors**  $b_k$  is **away from 0**, and in what direction (positive or negative).

- Since  $b_k$  is an **estimate** of the true (unknown) coefficient  $\beta_k$ :

- Large positive** values of  $t$  provide evidence in favor of  $H_a : \beta_k > 0$ .
- Large negative** values of  $t$  provide evidence in favor of  $H_a : \beta_k < 0$ .
- Both large positive and large negative** values of  $t$  provide evidence in favor of  $H_a : \beta_k \neq 0$ .

- Now suppose the **residuals\***  $e_1, e_2, \dots, e_n$  are a sample from a  $\mathbf{N}(0, \sigma)$  distribution or that  $n$  is **large**.

In this case, the **null distribution** is as follows.

\* More formally, the **errors**  $\epsilon_1, \epsilon_2, \dots, \epsilon_n$  in the regression **model**.

## Notes

---



---



---



---



---



---



---

## Notes

---



---



---



---



---



---



---

## Notes

---



---



---



---



---



---



---

## Notes

---



---



---



---



---



---



---

**Sampling Distribution of  $t$  Under  $H_0$ :** If  $t$  is the test statistic in a  $t$  test for a model coefficient, then when

$$H_0 : \beta_k = 0$$

is true,

$$t \sim t(n - (p + 1)).$$

- **P-values** and **rejection regions** are obtained from the appropriate tail(s) of the  $t(n - (p + 1))$  **distribution**.

- The  **$t$  test statistic** and **p-value** for the  **$t$  tests for the coefficients** are reported in the output of statistical software.
- The software also reports results of a  **$t$  test for the  $y$ -intercept**:

$$H_0 : \beta_0 = 0$$

$$H_a : \beta_0 \neq 0$$

but this is usually of little interest.

- The software summarizes the results in a **regression table** of the form below.

	Estimated Coefficient	Standard Error	$t$	P-value
Intercept	$b_0$	$S_{b_0}$	$t = b_0/S_{b_0}$	$p$
$X_1$	$b_1$	$S_{b_1}$	$t = b_1/S_{b_1}$	$p$
$X_2$	$b_2$	$S_{b_2}$	$t = b_2/S_{b_2}$	$p$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$X_p$	$b_p$	$S_{b_p}$	$t = b_p/S_{b_p}$	$p$

Notes

---

---

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

---

---

## Example

For the study of **water consumption**, **wealth**, and **population size** of  $n = 28$  U.S. cities, the  $t$  test results (obtained using software) are below.

	Estimated Coefficient	Standard Error	$t$	P-value
Intercept	6.48	0.139	46.51	0.000
Wealth	0.11	0.124	0.87	0.391
Population	0.16	0.026	6.12	0.000

Thus, the equation of the **fitted regression model** is

$$\hat{Y} = 6.48 + 0.11 X_1 + 0.16 X_2.$$

Notes

---

---

---

---

---

---

---

---

---

---

For **wealth**, the hypotheses are

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

The observed **test statistic** value is  $t = 0.87$  and the **p-value** is **0.391**.

Thus, using  $\alpha = 0.05$ , we **fail to reject**  $H_0$  and conclude that the observed **relationship** between **wealth** and **water consumption** is **not statistically significant** (*controlling for population size*).

Notes

---

---

---

---

---

---

---

---

---

---

For **population size**, the hypotheses are

$$H_0 : \beta_2 = 0$$

$$H_a : \beta_2 \neq 0$$

The observed **test statistic** value is  $t = 6.12$  and the **p-value** is **0.000**.

Thus we **reject**  $H_0$  and conclude that the observed **relationship** between **population** and **water consumption** is **statistically significant** (*controlling for wealth*).

Notes

---

---

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

---

---