INTERPRETATION OF THE MULTIPLE REGRESSION MODEL

- A dependent variable in the population may ***not only*** be influenced by one independent variable but by a whole set of independent variables.

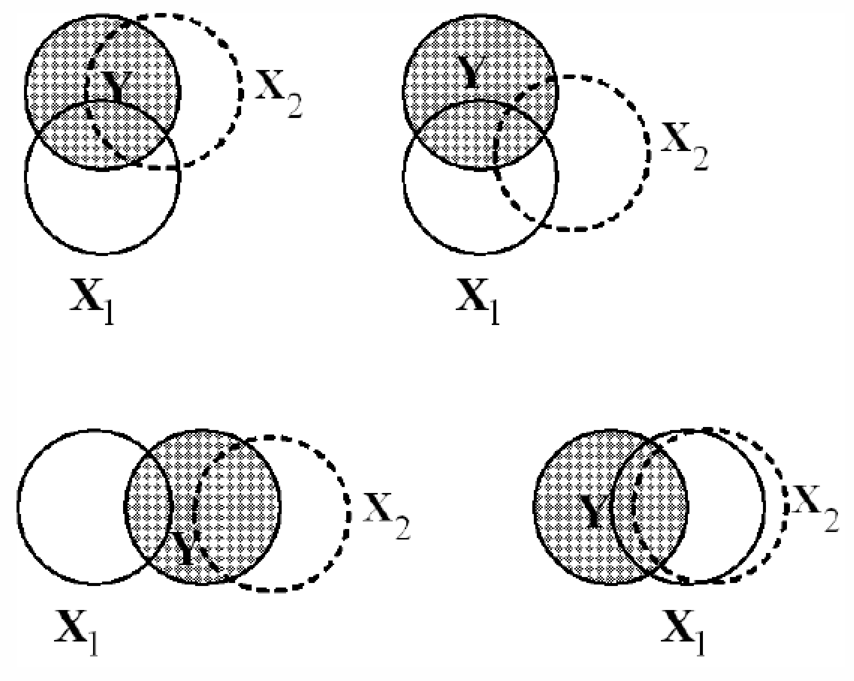
- These independent variables may ***not only*** be correlated with the dependent variable but also ***among each other***.

  $\Rightarrow$ This leads to the question how we can measure the individual effects (or contributions) of single variables?

- **Ballentine Venn Diagram**

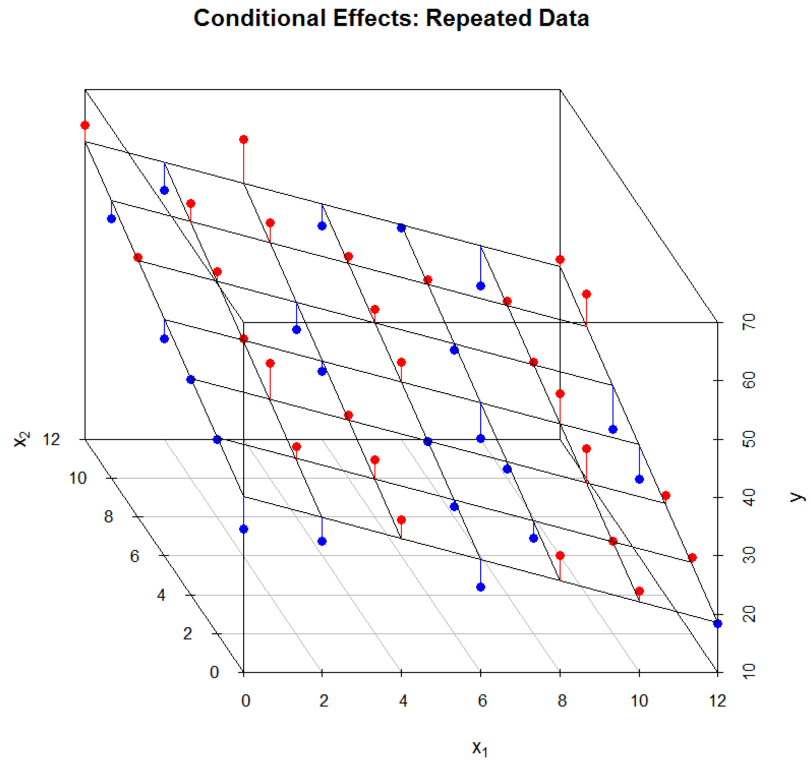
  Shared explained variance and correlation among independent variables

- Additional variables reduce the stochastic error in the dependent variable (equivalent to RSS).



MODEL WITH TWO INDEPENDENT VARIABLES

- Discussion of plot with two independent variables:
  - In multiple linear regression the independent variables erect a "***hyper***"-plane.
  - The ***slope*** of the surface along the axes X_1 and X_2 is given by the parameters b_1 and b_2. That is, the slope b_1 measures the variation of $\mathbf{y}$ with respect $\boldsymbol{x}_1$ at a given level of $\boldsymbol{x}_2$ (i.e., $x_{i2} = c$ held constant at a particular value c).
  - Analogue interpretation for slope b_2.
  - The regression parameters are also called ***partial regression coefficients*** because of the underlying assumption that if the remaining independent control variables are held constant at given levels, the ***relationship (slope) does not change*** from one level of the other independent control variables to another level.



- The intercept is given again by b_0 (here $x_{i1} = 0$ and $x_{i2} = 0$)
  That is, $E(Y_i \mid X_{i1} = 0, X_{i2} = 0) = \beta_0$ (see Ham p 66).

- How are X_1 and X_2 correlated in the given example?

PARTIAL REGRESSION COEFFICIENTS

- Experimental studies: We can control the correlations among the independent variables. $\Rightarrow$ The desired zero correlation level is achieved by assigning of the observations to different treatment level combinations. $\Rightarrow$ scatterplot of independent variable lacks a pattern

- Observational studies: We can only control for the spurious/confounding effects of other variables with a statistical approach. *Fortunately OLS does this for us!*

- In multivariate regression analysis the individual parameters express the relationship between the dependent and one independent variable ***while statistically holding the effects all other variables constant***.

- The set of estimated parameters $\{b_0, b_1, \ldots, b_{K-1}\}$ of the included variables may change as new variables are added to the model because of variables correlations among the independent variables.
  Only if the independent variables are all uncorrelated among each other, no change will occur.

- See example in Hamilton: *WaterUse81 = f(Income)* **versus** *WaterUse81 = f(Income, WaterUse80)*.

Table 2.2 Regression of 1981 water use on household income, with annotations

Source	SS	df	MS	
Model	190820566[1]	1[4]	190820566[7]	Number of obs = 496[10]
Residual	902418143[2]	494[5]	1826757.38[8]	$F(1, 494)$ = 104.46[11]
Total	1.0932e + 09[3]	495[6]	2208563.05[9]	Prob > F = 0.0000[12]
				R-square = 0.1745[13]
				Adj R-square = 0.1729[14]
				Root MSE = 1351.6[15]

Variable	Coefficient	Std. Error	t	Prob > |t|	Mean
water81					2298.387[24]
income	47.54869[16]	4.652286[18]	10.221[20]	0.000[22]	23.07661[25]
_cons	1201.124[17]	123.3245[19]	9.740[21]	0.000[23]	1

[1] explained sum of squares, ESS (equation [2.13])

Table 3.1 Regression of postshortage (1981) water use on income and preshortage (1980) water use.

Source	SS	df	MS	
Model	671025350	2	335512675	Number of obs = 496
Residual	422213359	493	856416.551	$F(2, 493)$ = 391.76
Total	1.0932e + 09	495	2208563.05	Prob > F = 0.0000
				R-square = 0.6138
				Adj R-square = 0.6122
				Root MSE = 925.43

Variable	Coefficient	Std. Error	t	Prob > |t|	Mean
water81					2298.387
income	20.54504	3.38341	6.072	0.000	23.07661
water80	.5931267	.0250482	23.679	0.000	2732.056
_cons	203.8217	94.36129	2.160	0.031	1

Questions:

[1] Is the estimated coefficient in the bivariate model biased or does it reflect the value of its underlying population?

[2] What does the previous water consumption measure?

[3] Are the regression assumptions still satisfied by using a temporarily lagged dependent variable on the right-hand side of the equation?

PARTIAL EFFECTS AND LEVERAGE PLOT

- Interpretation of regression residuals:
  - The regression residuals $e_i = y_i - \hat{y}_i$ are free from any linear effect with the model's independent variables.
  - They measure the remaining (unexplained) variation of the dependent variable y after ***accounting for*** the <u>included independent variables</u>.
- Therefore, the residuals are uncorrelated, or more generally "orthogonal", with the independent variables, i.e., $\sum_{i=1}^{n} x_{ij} \cdot e_i = 0$.

  Since the constant unity vector $\mathbf{1} = (1,1,\ldots,1)^T$ is part of the independent variables, they also sum to zero, that is, $\sum_{i=1}^{n} 1 \cdot e_i = 0$.

  $\Rightarrow$ This provides a justification for keeping even an insignificant intercept in the model because otherwise the residuals may not sum to zero anymore.

  A proof of these properties is easily accomplished with matrix algebra.
- Work through the water usage example **HAM** pp 70-71:

  (1) Removing effect of X_2 (water80)

$$\hat{Y}_i = 203.8 + 20.5X_{i1} + 0.59X_{i2}$$

– – – – –

$$Y_i = 537.9 + 0.64X_{i2} + e_{i,Y|X_2}$$

$$X_{i1} = 16.26 + 0.0025X_{i2} + e_{i,X_1|X_2}$$

$$\hat{e}_{i,Y|X_2} = 0 + 20.5e_{i,X_1|X_2}$$

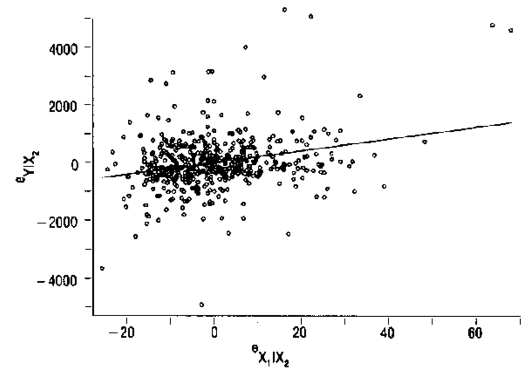


Figure 3.1 Partial regression leverage plot: postshortage water use (Y) versus income (X_1), adjusting for preshortage water use (X_2).

Questions:

[a] Why is the b_0 coefficient zero? Answer: the sum of the residuals is zero, thus the means of $e_{X_1|X_2}$ and $e_{Y|X_2}$ lies at the origin (0,0) of the coordinate system.

[b] What happens to the residuals $e_{X_1|X_2}$ if X_1 and X_2 are perfectly correlated? What is the impact on the partial regression equation?

[c] What happens if X_1 and X_2 are uncorrelated?

(2) Removing the effect of X_1 (Income)

$$Y_i = 1201.1 + 47.5 X_{i1} + e_{i,Y|X_1}$$
$$X_{i2} = 1681.4 + 45.5 X_{i1} + e_{i,X_2|X_1}$$
$$\hat{e}_{i,Y|X_1} = 0 + 0.59 e_{i,X_2|X_1}$$

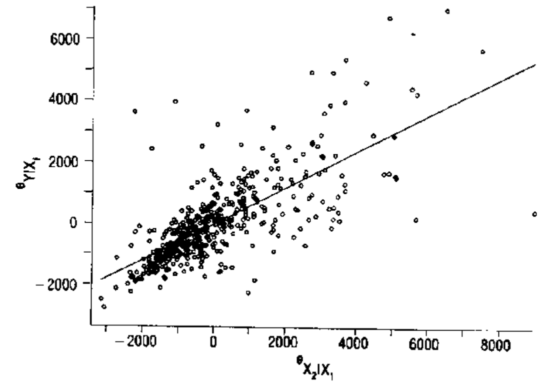


Figure 3.2 Partial regression leverage plot: postshortage water use (Y) versus preshortage water use (X_2), adjusting for income (X_1).

- Compared to the sign of the regression coefficient in the bivariate model, it is possible for a multiple model that the sign of a regression parameter ***changes***, that is, the direction of a partial parameter changes.

- **Partial correlation:** It is the correlation between *Y* and X_1 after the linear effects of $\{X_2, \ldots X_{K-1}\}$ has been removed $Corr(e_{Y|X_2,\ldots X_{K-1}}, e_{X_1|X_2,\ldots X_{K-1}})$

Selection Criteria for the Independent Variables

- A theory should guide us in our decisions, which variables have a ***meaningful influence*** on the endogenous variable (they are significant) and in which ***direction*** (sign of the partial regression coefficient) this influence points.

However, in many cases theory is not sufficiently explicit and we may also search for a set of relevant variables ($\Rightarrow$ exploratory regression analysis).

- **Adding additional variables:**
  - R^2 always increases (or stays at least the same) but not necessarily R^2_{adj} (Why?)
  - Important variables => their estimated coefficient is significantly different from zero
  - Coefficients of spurious variables may shrink.
- **General goal:** balance between ***simplicity*** and ***complexity*** of the regression model (***statistical concept of parsimony***): We want to describe the variability in the dependent variable as efficiently as possible.
  - R^2_{adj} reflects this concept by penalizing adding additional independent variables.
  - Akaike's information criterion, $AIC = -\log(likelihood) + 2 \cdot (K-1)$, also makes use of this concept. <u>Smaller AIC values are preferred</u>:
    $-\log(likelihood) \downarrow \text{when } K \uparrow$, that is, each additional estimated parameter increases *AIC* by 2, therefore the $-\log(likelihood)$ needs to shrink by at least by 2.

- Note: If we include $n-1$ independent randomly generated variables we would obtain a perfectly fitting model, i.e., $R^2 = 1$, even though none of these random variables is relevant. However, such a model is as ***complex*** (n estimated parameters based on n observations) as our original data and therefore violates the ***paradigm of parsimony***.

- Consequences of a misspecified regression model:

  - ***Including irrelevant variables****:* Coefficient statistical close to zero. R^2 increases only marginally. Standard errors of all estimated parameters will increases (=> What are the consequences). Additional irrelevant variables increase unnecessarily the complexity of the model.

  - ***Omitted relevant variables****:* If the omitted variables are correlated with other variables in the model, then the parameter estimates of the included variables becomes biased, i.e., $E(b_k) \neq \beta_k$. Unrealistic simple model. Standard errors of parameters in the model usually smaller.
    General aim: Minimize the ***mean square error*** of the included parameter estimates.

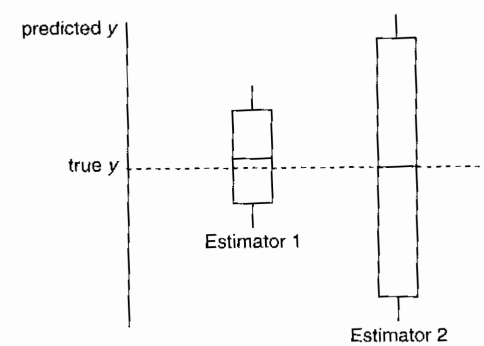


Figure 2.21 *Bias and variance in estimators*

$$MSE = E\left[b_k - \beta_k\right]^2 \qquad Var\left[b_k\right] = E\left[b_k - E\left(b_k\right)\right]^2$$
$$= Var\left[b_k\right] + bias^2 \qquad bias^2 = \left[E\left(b_k\right) - \beta_k\right]^2$$

Tradeoff between standard error and bias. We may be willing to accept a small bias if the variance of the estimated parameter is decreased substantially.

Discussion Seven Variables Example (HAM p 74)

- State a hypothesis about each included variable.
- Compare against model with only **Income** and **Water80**. Which parameters are stable?
- Status of the **Retire** variable. => binary 0/1 indicator variable with 0=*not retired* and 1=*retired*
- Order of partial regression coefficients in terms of *t*-values

Table 3.2 Regression of postshortage water use on income, preshortage water use, education, retirement, number of people resident, and increase in people resident

Source	SS	df	MS	
Model	740477522	6	123412920	Number of obs = 496
Residual	352761188	489	721393.022	$F(6, 489)$ = 171.08
				Prob > F = 0.0000
				R-square = 0.6773
Total	1.0932e + 09	495	2208563.05	Adj R-square = 0.6734
				Root MSE = 849.35

Variable	Coefficient	Std. Error	t	Prob > |t|	Mean
water81					2298.387
income	20.96699	3.463719	6.053	0.000	23.07661
water80	.49194	.0263478	18.671	0.000	2732.056
educat	−41.86552	13.22031	−3.167	0.002	14.00403
retire	189.1843	95.02142	1.991	0.047	.2943548
peop81	248.197	28.7248	8.641	0.000	3.072581
cpeop	96.4536	80.51903	1.198	0.232	−.0383065
_cons	242.2204	206.8638	1.171	0.242	1

- What does the change in the number of people measure?
- Interpretation: water consumption increases if …, and it decreases if…

STANDARDIZED REGRESSION COEFFICIENTS (BETA COEFFICIENTS)

- Same interpretation as in bivariate regression analysis but now in terms of ***variations of standard deviations*** and not in terms of original variable measurement units: $beta_i = \frac{s_{x_i} \cdot b_i}{s_y}$

  That is, if x_i changes by one standard deviation, how many standard deviations does y change?

- The intercept is always zero because regression goes through the origin, i.e., the mean of standardized variables is always zero.
- The larger the absolute value (maximum value is less than $|1|$) the more influence has the independent variable on the variation of the dependent variable.
- This allows to compare the importance of individual variable in one fixed model:

Variable	Beta-weight
Income	0.18
water80	0.58

Educat	-0.09
Retire	0.06
peop81	0.28
Cpeop	0.03

- Beta coefficients are not comparable among different samples or investigations because the variance of the dependent and independent variables may vary slightly from sample to sample.
- In contrast to bivariate regression analysis, the ***beta coefficients are no longer correlations*** between the dependent and the independent variables.

GLOBAL *F*-TEST

- Null hypothesis testing for ***all parameters*** (except intercept) equal to zero.
- ***Global*** (or ***omnibus***) *F*-test:
  $H_0: \beta_1 = \beta_2 = \ldots = \beta_{K-1} = 0$ against
  $H_1: \beta_j \neq 0$ for at least on $j \in \{1, \ldots, K-1\}$
- The test statistic is: $F_{df_1, df_2} = \dfrac{ESS/(K-1)}{RSS/(n-K)}$ where $df_1 = K-1$ and $df_2 = n-K$

PARTIAL *F*-TEST

- **Nested models**: The ***ordinate*** model has more independent variables (say *K*) than the ***subordinate*** model, which has *H* ***fewer independent*** variables with $0 < H < K$.
- The subordinate model consists of a ***sub-set*** of variables from the ordinate model.
- This allows a comparison across models. If models are ***not nested*** a direct comparison is impossible.

- To test whether these *H* independent variables add significantly to the model the *partial F*-test can be used with

$$F_{n-K}^{H} = \frac{\left(RSS_{K-H} - RSS_{K}\right)/H}{RSS_{K}/n-K}$$

with RSS_{K-H} is the *restricted* model with H variables less and RSS_K is the full model.
Note that $RSS_K \leq RSS_{K-H}$, because the residual sum of squares decreases (or stays the same) as additional variables are included into the model.

- Compare model without ***Income*** and ***Education*** against full model by using the tabulated data. I.e: $H_0 : \beta_1 = \beta_3 = 0$. See **HAM** p 81.

Table 3.3 Regression of postshortage water use omitting income and education

Source	SS	df	MS	
Model	712718346	4	178179587	Number of obs = 496
Residual	380520363	491	774990.557	$F(4, 491)$ = 229.91
Total	1.0932*e* + 09	495	2208563.05	Prob > F = 0.0000
				R-square = 0.6519
				Adj R-square = 0.6491
				Root MSE = 880.34

Table 3.2 Regression of postshortage water use on income, preshortage water use, education, retirement, number of people resident, and increase in people resident

Source	SS	df	MS	
Model	740477522	6	123412920	Number of obs = 496
Residual	352761188	489	721393.022	$F(6, 489)$ = 171.08
Total	1.0932*e* + 09	495	2208563.05	Prob > F = 0.0000
				R-square = 0.6773
				Adj R-square = 0.6734
				Root MSE = 849.35

$$F_{489}^{3} = \frac{\left(RSS\{5\} - RSS(7)\right)/2}{RSS(7)/(496-7)}$$

$$= \frac{\left(380,520,363 - 352,761,188\right)/2}{352,761,188/489}$$

$$= 19.24$$

- The global *F*-test as special case of the partial *F*-test (**HAM** eq 3.29): I.e. $H_0 : \beta_1 = \ldots = \beta_{K-1} = 0$. Here $H = K - 1$ and the constant vector remains the only variable in the restricted regression equation.

$$
\begin{aligned}
F_{n-K}^{K-1} &= \frac{(RSS_1 - RSS_K)/(K-1)}{RSS_K/(n-K)} \\
&= \frac{(TSS_Y - RSS_K)/(K-1)}{RSS_K/(n-K)} \\
&= \frac{ESS/(K-1)}{RSS_K/(n-K)}
\end{aligned}
$$

- The single parameter *t*-test is a special case of the partial *F*-test with the null hypothesis $H_0 : \beta_k = 0$ and the number of excluded parameters being $H = 1$.
  However, since the *F*-statistic is always positive a directed hypothesis cannot be tested.

EXCURSION: R'S ANOVA-FUNCTIONS

- The **anova** function can perform partial *F*-tests for nested models. ⇒ **`anova(lm.mod1,lm.mod2)`**.

- Applied on a single model, e.g., `lm(y~x1+x2+x3)`, it performs a sequence of nested tests and provides the residual sum of squares of the full model. The sequence of test are:
  (i) `lm(y~1)` against `lm(y~x1)`,
  (ii) `lm(y~x1)` against `lm(y~x1+x2)`, and
  (iii) `lm(y~x1+x2)` against `lm(y~x1+x2+x3)`
  In general, only the last comparison is meaningful unless one has a clear expectation of the order of the independent variables.
  This is known as *SAS type I ANOVA*.

- In contrast, the function **Anova** in the **car** library tests an alternative set of hypotheses. Its sequence of tests is
  (i) `lm(y~x2+x3)` against `lm(y~x1+x2+x3)`,
  (ii) `lm(y~x1+x3)` against `lm(y~x1+x2+x3)`, and
  (iii) `lm(y~x1+x2)` against `lm(y~x1+x2+x3)`
  This is known as *SAS type II ANOVA*.

- For more details see Fox&Weisberg (2019), pages 260-264.