Chapter 6 Multiple Regression

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I

6.1 Multiple regression models

We now add more predictors, linearly, to the model. For example let's add one more to the simple linear regression model:

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

with the usual $E(\epsilon_i) = 0$. For *any* Y in this population with predictors (x_1, x_2) we have

$$\mu(x_1, x_2) = E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

The triple $(x_1, x_2, \mu(x_1, x_2)) = (x_1, x_2, \beta_0 + \beta_1 x_1 + \beta_2 x_2)$ describes a plane in $\mathbb{R}^3$ (p. 215).

Multiple regression models

Generally, for $k = p - 1$ predictors $x_1, \ldots, x_k$ our model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i, \tag{6.7}$$

with mean

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}. \tag{6.8}$$

- β_0 is mean response when all predictors equal zero (if this makes sense).
- β_j is the change in mean response when x_j is increased by one unit *but the remaining predictors are held constant*.
- We will assume normal errors:

$$\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} N(0, \sigma^2).$$

Dwayne Portrait Studio data (Section 6.9)

Dwayne Portrait Studio is doing a sales analysis based on data from 21 cities.

- Y = sales (thousands of dollars) for a city
- x_1 = number of people 16 years or younger (thousands)
- x_2 = per capita disposable income (thousands of dollars)

Assume the linear model is appropriate. One way to check marginal relationships is through a scatterplot matrix. However, these are not infallible.

For these data, is β_0 interpretable?

β_2 is the change in the mean response for a thousand-dollar increase in disposable income, holding "number of people under 16 years old" constant.

SAS code

```
data studio;
 input people16 income sales @@;
 label people16='Number 16 and under (thousands)'
       income ='Per capita disposable income ($1000)'
       sales  ='Sales ($1000$)';
datalines;
  68.5  16.7  174.4  45.2  16.8  164.4  91.3  18.2  244.2  47.8  16.3  154.6
  46.9  17.3  181.6  66.1  18.2  207.5  49.5  15.9  152.8  52.0  17.2  163.2
  48.9  16.6  145.4  38.4  16.0  137.2  87.9  18.3  241.9  72.8  17.1  191.1
  88.4  17.4  232.0  42.9  15.8  145.3  52.5  17.8  161.1  85.7  18.4  209.7
  41.3  16.5  146.4  51.7  16.3  144.0  89.6  18.1  232.6  82.7  19.1  224.1
  52.3  16.0  166.5
;

proc sgscatter; matrix people16 income sales / diagonal=(histogram kernel); run;

options nocenter;
proc reg data=studio;
 model sales=people16 income / clb; * clb gives 95% CI for betas;
run;                                * alpha=0.9 for 90% CI, etc.;
```

SAS output

The REG Procedure

Analysis of Variance

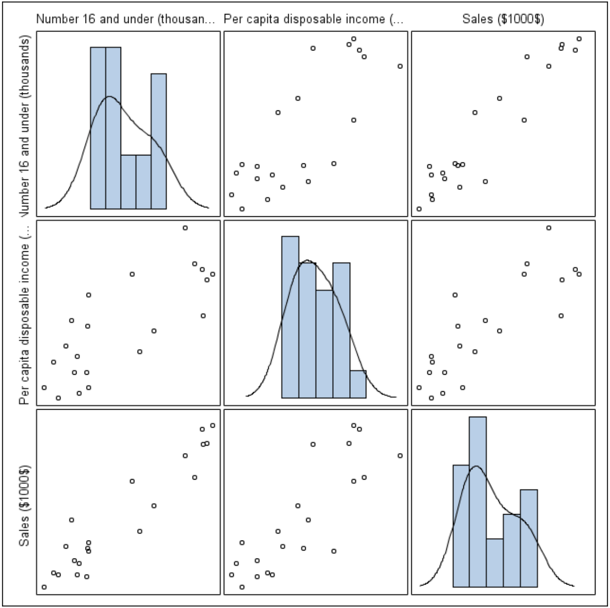
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	24015	12008	99.10	<.0001
Error	18	2180.92741	121.16263		
Corrected Total	20	26196			

Root MSE	11.00739	R-Square	0.9167
Dependent Mean	181.90476	Adj R-Sq	0.9075
Coeff Var	6.05118		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	-68.85707	60.01695	-1.15	0.2663	-194.94801	57.23387
people16	Number 16 and under (thousands)	1	1.45456	0.21178	6.87	<.0001	1.00962	1.89950
income	Per capita disposable income ($1000)	1	9.36550	4.06396	2.30	0.0333	0.82744	17.90356

Scatterplot matrix



The general linear model encompasses...

Qualitative predictors

Example: Dichotomous predictor

- Y = length of hospital stay
- x_1 = gender of patient ($x_1 = 0$ male, $x_1 = 1$ female)
- x_2 = severity of disease on 100 point scale

$$E(Y) = \left\{ \begin{array}{ll} \beta_0 + \beta_2 x_2 & \text{males} \\ \beta_0 + \beta_1 + \beta_2 x_2 & \text{females} \end{array} \right\}.$$

Response functions are two parallel lines, shifted by β_1 units...so-called “ANCOVA” model.

The general linear model encompasses...

Polynomial regression
Often appropriate for curvilinear relationships between response and predictor.
Example:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon.$$

Letting $x_2 = x_1^2$ this is in the form of the general linear model.

Transformed response
Example:

$$\log Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon.$$

Let $Y^* = \log(Y)$ and get general linear model.

The general linear model encompasses...

Interaction effects
Example:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon.$$

Let $x_3 = x_1 x_2$ and get general linear model.

Key: All of these models are *linear in the coefficients*, the β_j terms. An example of a model that is *not* in general linear model form is exponential growth:

$$Y = \beta_0 \exp(\beta_1 x) + \epsilon.$$

6.2 General linear model in matrix terms

Let $\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$ be the *response vector*.

Let $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$ be the *design matrix*

containing the predictor variables. The first column is a place-holder for the intercept term. What does each column represent? What does each row represent?

Let $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$ be the unkown vector of *regression coefficients*.

Let $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$ be the unobserved *error vector*.

General linear model in matrix terms

The general linear model is written in matrix terms as

$$\underbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}}_{n \times 1} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}}_{n \times p} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}}_{p \times 1} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}}_{n \times 1},$$

where $p = k + 1$, or succinctly as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

General linear model in matrix terms

Minimal assumptions about the random error vector $\boldsymbol{\epsilon}$ are

$$E(\boldsymbol{\epsilon}) = \mathbf{0} \text{ and } \text{cov}(\boldsymbol{\epsilon}) = \mathbf{I}_n\sigma^2,$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix (zero except for 1's along the diagonal).

In general, we will go farther and assume

$$\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \mathbf{I}_n\sigma^2).$$

This allows use to construct t and F tests, obtain confidence intervals, etc.

Writing the model like this saves a *lot* of time and space as we go along.

6.3 Fitting the model

Estimating $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$

Recall least-squares method: minimize

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_{i1} + \cdots \beta_k x_{ik})]^2,$$

as a function of $\boldsymbol{\beta}$. Vector calculus can show that the least-squares estimates are

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

typically found using a computer package. **Note**: there is a typo in the book (equation (6.25) p. 223).

6.4 Fitted values & residuals

The *fitted values* are in the vector

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \mathbf{Xb} = \underbrace{[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']}_{\text{projection matrix}} \mathbf{Y} = \mathbf{HY}. \tag{6.30}$$

The *residuals* are in the vector

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{Xb} = \underbrace{[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']}_{\text{projection matrix}} \mathbf{Y}. \tag{6.31}$$

Hat matrix & Dwayne Studios

$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is called the "hat matrix." We'll use it shortly when we talk about diagnostics. Note also that $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$.

Back to **Dwayne Portrait Studio data**. From SAS,

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} -68.857 \\ 1.455 \\ 9.366 \end{bmatrix},$$

so the fitted regression line is

$$\hat{Y} = -68.857 + 1.455x_1 + 9.366x_2.$$

Interpretation

- **Interpretation of** b_1: We *estimate* that for 1000 person increase in persons 16 and under, mean sales increase by \$1,455 (1.455 thousand dollars) holding per capita disposable income constant.
- **Interpretation of** b_2: We *estimate* that for each \$1000 increase in per capita disposable income, mean sales increase by \$9,366, holding the number of people under 16 constant.

6.5 Analysis of variance

Again, in multiple regression we can decompose the total sum of squares into the SSR and SSE pieces. The table is now

Source	SS	df	MS	$E(MS)$
Regression	SSR$= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	$p-1$	$\frac{SSR}{p-1}$	$\sigma^2 +$ QF
Error	SSE$= \sum_{i=1}^{n}(Y_i - \hat{Y})^2$	$n-p$	$\frac{SSE}{n-p}$	σ^2
Total	SSTO$= \sum_{i=1}^{n}(Y_i - \bar{Y})^2$	$n-1$		

where $p = k + 1$.
Here, QF stands for "quadratic form" and is given by

$$\mathrm{QF} = \tfrac{1}{2}\sum_{j=1}^{k}\sum_{s=1}^{k}\beta_j\beta_s\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{is} - \bar{x}_s) \geq 0.$$

Note that QF$= 0 \Leftrightarrow \beta_1 = \beta_2 = \cdots = \beta_k = 0$.

Overall F-test for a regression relationship (p. 226)

In multiple regression, our F-test based on $F^* = \frac{MSR}{MSE}$ tests whether the *entire set* of predictors $x_1, \dots, x_k$ explains a significant amount of the variation in Y.

If $MSR \approx MSE$, there's no evidence that *any* of the predictors are useful. If $MSR >> MSE$, then some or all of them are useful.

Formally, the F-test tests $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ versus H_a : at least one of these is not zero. If $F^* > F_{p-1,n-p}(1-\alpha)$, we reject H_0 and conclude that *something* is going on, there is *some* relationship between or more of the $x_1, \dots, x_k$ and Y. SAS provides a p-value for this test.

R^2 is how much variability soaked up by model

The coefficient of multiple deterimation is

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \tag{6.40}$$

measures the proportion of sample variation in Y explained by its *linear* relationship with the predictors $x_1, \ldots, x_k$. As before, $0 \leq R^2 \leq 1$.

When we add a predictor to the model R^2 can only increase.

The adjusted R^2

$$R_a^2 = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)} \tag{6.42}$$

accounts for the number of predictors in the model. It may decrease when we add useless predictors to the model.

Dwayne Studios, ANOVA table, R^2, & R_a^2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	24015	12008	99.10	<.0001
Error	18	2180.92741	121.16263		
Corrected Total	20	26196			

Root MSE	11.00739	R-Square	0.9167
Dependent Mean	181.90476	Adj R-Sq	0.9075
Coeff Var	6.05118		

We reject $H_0 : \beta_1 = \beta_2 = 0$ at any reasonable significance level α. About 92% of the total variability in the data is explained by the linear regression model.

Inference about individual regression parameters

The overall F-test concerns the *entire set* of predictors $x_1, \ldots, x_k$.

If the F-test is significant (if we reject H_0), we will want to determine *which* of the individual predictors contribute significantly to the model.

We will talk about this shortly, but the main methods are forward selection, backwards elimination, stepwise procedures, C_p, and R_a^2.

Aside: There are also *fancy* new methods including LASSO, LARS, etc. These are used when there's *lots* of predictors, e.g. $p = 500$ or $p = 20,000$.

Recall: If $\mathbf{Y}$ is a random vector, then its *expected value* is also a vector

$$E(\mathbf{Y}) = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{bmatrix}.$$

The random vector $\mathbf{Y}$ also has a *covariance matrix*

$$\text{cov}(\mathbf{Y}) = \begin{bmatrix} \text{cov}(Y_1, Y_1) & \text{cov}(Y_1, Y_2) & \cdots & \text{cov}(Y_1, Y_n) \\ \text{cov}(Y_2, Y_1) & \text{cov}(Y_2, Y_2) & \cdots & \text{cov}(Y_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(Y_n, Y_1) & \text{cov}(Y_n, Y_2) & \cdots & \text{cov}(Y_n, Y_n) \end{bmatrix}.$$

Multivariate normal

An aside: the *multivariate normal* density is given by

$$f(\mathbf{y}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\{-0.5(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\},$$

where $\mathbf{y} \in \mathbb{R}^d$. We write

$$\mathbf{Y} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Then $E(\mathbf{Y}) = \boldsymbol{\mu}$ and $\mathrm{cov}(\mathbf{Y}) = \boldsymbol{\Sigma}$.

For the general linear model,

$$\mathbf{Y} \sim N_n(\mathbf{X}\beta, \sigma^2\mathbf{I}_n).$$

Error vector

Note that along the diagonal of $\text{cov}(\mathbf{Y})$, $\text{cov}(Y_i, Y_i) = \text{var}(Y_i)$.

For the general linear model,

$$E(\boldsymbol{\epsilon}) = \mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

$$\text{cov}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}.$$

$$\text{cov}(\mathbf{Y}) = \text{cov}(\underbrace{\mathbf{X}\boldsymbol{\beta}}_{\text{constant}} + \underbrace{\boldsymbol{\epsilon}}_{\text{random}}) = \text{cov}(\boldsymbol{\epsilon}) = \mathbf{I}_n\sigma^2.$$

Fact: If $\mathbf{A}$ is a constant matrix, $\mathbf{a}$ is a constant vector, and $\mathbf{Y}$ is any random vector, then

$$E(\mathbf{AY} + \mathbf{a}) = \mathbf{A}E(\mathbf{Y}) + \mathbf{a},$$

$$\text{cov}(\mathbf{AY} + \mathbf{a}) = \mathbf{A}\text{cov}(\mathbf{Y})\mathbf{A}'.$$

Back to the general linear model

For $\hat{\mathbf{Y}} = \mathbf{HY}$,

$$E(\hat{\mathbf{Y}}) = \mathbf{H}E(\mathbf{Y}) = \mathbf{HX}\boldsymbol{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}.$$

$$\mathrm{cov}(\hat{\mathbf{Y}}) = \mathbf{H}\mathrm{cov}(\mathbf{Y})\mathbf{H}' = \sigma^2\mathbf{H},$$

since $\mathbf{HH}' = \mathbf{H}$ (property of a *projection matrix*).
For $\mathbf{e} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$,

$$E(\mathbf{e}) = (\mathbf{I}_n - \mathbf{H})E(\mathbf{Y}) = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} - \mathbf{HX}\boldsymbol{\beta} = \mathbf{0},$$

as $\mathbf{HX} = \mathbf{X}$ (projection matrix again).

$$\mathrm{cov}(\mathbf{e}) = (\mathbf{I}_n - \mathbf{H})\mathrm{cov}(\mathbf{Y})(\mathbf{I}_n - \mathbf{H})' = \sigma^2(\mathbf{I}_n - \mathbf{H}).$$

(Guess why...)

Finally, $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is *unbiased*

$$E(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta},$$

and has covariance matrix

$$\begin{aligned}
\text{cov}(\mathbf{b}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{cov}(\mathbf{Y})[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.
\end{aligned}$$

Presto!

$$\mathbf{b} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$

Table of regression effects (p. 228)

From the previous slide, the jth estimated coefficient β_j,

$$\mathrm{var}(b_j) = \sigma^2 c_{jj},$$

where c_{jj} is the jth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. *Estimate* the standard deviation of b_j by its standard error $\mathrm{se}(b_j) = \sqrt{MSEc_{jj}}$ yielding

$$\frac{b_j - \beta_j}{se(b_j)} \sim t_{n-p} \tag{6.49}$$

Note: SAS gives each $\mathrm{se}(b_j)$ as well as b_j, $t_j^* = b_j/se(b_j)$, and a p-value for testing each $H_0 : \beta_j = 0$.

Dwayne output

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	-68.85707	60.01695	-1.15	0.2663	-194.94801	57.23387
people16	Number 16 and under (thousands)	1	1.45456	0.21178	6.87	<.0001	1.00962	1.89950
income	Per capita disposable income ($1000)	1	9.36550	4.06396	2.30	0.0333	0.82744	17.90356

We reject $H_0 : \beta_1 = 0$ at the $\alpha = 0.01$ level and $\beta_2 = 0$ at the $\alpha = 0.05$ level.

Individual tests of $H_0 : \beta_j = 0$

Note: A test of $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$ – available in the table of regression coefficients – is a test of whether predictor x_j is necessary in a model *with the other remaining predictors included*. For the Studio Data:

- The SAS summary gives us $F^* = MSR/MSE = 99.10$ with associated p-value < 0.0001 (it is actually 2×10^{-10}!). We strongly reject (at any reasonable α) $H_0 : \beta_1 = \beta_2 = 0$.
- 95% CI's are $(1.01, 1.90)$ for β_1 and $(0.83, 17.90)$ for β_2.
- For example, we are 95% confident that mean sales increases by \$1010 to \$1900 for every 1000 increase in kids 16 and under, holding income constant.
- For $H_0 : \beta_1 = 0$ we get $p < 0.0001$; for $H_0 : \beta_2 = 0$ we get $p = 0.03$. Are people under 16, x_1, and income, x_2, important in the model?

Optional output

Model statement `model sales=people16 income / covb corrb;`
gives $\widehat{cov}(\mathbf{b})$ and $\widehat{corr}(\mathbf{b})$

Covariance of Estimates

Variable	Label	Intercept	people16	income
Intercept	Intercept	3602.0346743	8.7459395806	-241.4229923
people16	Number 16 and under (thousands)	8.7459395806	0.0448515096	-0.672442604
income	Per capita disposable income ($1000)	-241.4229923	-0.672442604	16.515755794

Correlation of Estimates

Variable	Label	Intercept	people16	income
Intercept	Intercept	1.0000	0.6881	-0.9898
people16	Number 16 and under (thousands)	0.6881	1.0000	-0.7813
income	Per capita disposable income ($1000)	-0.9898	-0.7813	1.0000

These are not typically used except for more "unusual" analyses; e.g. inference for β_1/β_2.