Chapter 6: Statistical Inference for Regression

6.1 Simple Regression Model

Population regression model:

$$Y_i = \alpha + \beta X_i + \varepsilon_i,\ i = 1, \ldots, n$$

$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

The errors ε_i represent

- omitted explanatory variables
- measurement error
- "inherently random component of Y"

Here we address *key assumptions* in order to do *inference* (confidence intervals and tests). But first more on …

Notation: From now on the text notation is mostly consistent in using capital letters for random variables and lowercase letters to denote observations, as in the first line above (I wrote it with a small "y" before). However, it is still somewhat variable as the regression model is variously written with capital and lowercase x on the right-hand side. A footnote on p. 100 explains that Fox uses lowercase x to stress that the value of x_i is fixed, either literally in an experimental design, or by conditioning on the observed value value x_i of X_i. So you see that the explanatory variable could be considered as a random variable or fixed numbers. In addition, chap 9 chooses to denote a vector of random variables by a lower case bold $\mathbf{y}$ as shown above, rather than a bold capital letter.

I tend to use the "hat" notation for estimates where as Fox uses latin letters for estimates of (greek letter) parameters.

Key assumptions in order to do *inference* (confidence intervals and tests):

- Linearity: $E(y_i) = \alpha + \beta x_i$,

  or $E(\varepsilon_i) = E(\varepsilon_i \mid x_i) = 0$
- Constant variance: $\mathrm{Var}(\varepsilon_i \mid x_i) = \mathrm{E}(Y_i - \alpha - \beta x_i)^2 = \sigma_\varepsilon^2$
- Normality: $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$
  - Question: Do we assume that the Y_i are normally distributed?
- Independence: random errors ε_i and ε_j are statistically (probabilistically) independent. This assumption is generally justified by knowledge of how the data were sampled/collected. For example, observations which represent
  - Random sample of subjects from large population, vs.
  - Time series, vs
  - Sample of children from a school with multiple classrooms. But here, the assumption will depend on what you are measuring. An analysis of `height ~ age`, independence between children may be a reasonable assumption, for for an analysis of `spelling ~ age`, probably not.
- X's are either
  - Fixed by design, or
  - Sampled/measured with error independent of ε_i

Examples for discussion:

- Average income -> Prestige for Canadian occupations
- Education -> Hourly wage rate for Canadian employees
- GDP per capita -> Infant mortality in the UN data base

6.1.2 Properties of least squares estimates:

$$\hat{\alpha} = \overline{Y} - B\overline{X}$$

$$\hat{\beta} = \frac{\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum\left(X_i - \overline{X}\right)^2}$$

or, in matrix notation

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

(1) L.S. estimates are linear in Y

$$\hat{\beta} = \sum m_i Y_i, \; m_i = \frac{x_i - \overline{x}}{\sum (x_j - \overline{x})^2}$$

$$\hat{\alpha} = ?\ (exercise)$$

(2) Sampling variance is

$$\begin{aligned}\mathrm{Var}\left(\hat{\beta}\right) &= \mathrm{Var}\left(\sum m_i Y_i\right) = \sum m_i^2 \,\mathrm{Var}(Y_i) + 2\sum_{i<j} m_i m_j \mathrm{Cov}(Y_i, Y_j) \\ &= \frac{1}{\left(\sum (x_j - \overline{x})^2\right)^2} \sum (x_i - \overline{x})^2 \sigma_\varepsilon^2 \\ &= \frac{\sigma_\varepsilon^2}{\sum (x_j - \overline{x})^2} = \frac{\sigma_\varepsilon^2}{(n-1)s_x^2}\end{aligned}$$

Question: If you are designing a study, how can you make the uncertainty (standard error) of $\hat{\beta}$ small?

Similarly, you can show

$$\begin{aligned}\mathrm{Var}\left(\hat{\alpha}\right) &= \mathrm{Var}\left(\sum q_i Y_i\right) \\ &= \frac{\sigma_\varepsilon^2 \sum x_i^2}{n\sum (x_j - \overline{x})^2}\end{aligned}$$

which is small(est) when the x's are centered at zero.

(3) If $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, then

$$\hat{\alpha} \sim N\left(\alpha, \frac{\sigma_\varepsilon^2 \sum x_i^2}{n\sum (x_j - \overline{x})^2}\right)$$

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma_\varepsilon^2}{\sum (x_j - \overline{x})^2}\right)$$

But what if the errors cannot be assumed normally distributed?

Recall that $\hat{\beta} = \sum m_i Y_i$, so that by the Central Limit Theorem $\hat{\beta}$ should be approximately large

(4) By the Gauss-Markov Theorem, of all linear estimators $\beta^* = \sum c_i Y_i$ that are unbiased, $\mathrm{E}\beta^* = \beta$, $\hat{\beta}_{LS}$ has the minimum variance --- it is the "most efficient" estimator.

(5) Under all the assumptions stated above, wich can be combined into one statement,

$$\varepsilon_i \underset{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

$(\hat{\alpha}, \hat{\beta})$ are *maximum likelihood* estimates. (Consider reading Appendix D (online) to the text by Fox, although this section goes into more detail than you need)

6.1.3 Confidence intervals and hypothesis tests:

$\mathrm{Var}\left(\hat{\beta}\right) = \dfrac{\sigma_\varepsilon^2}{(n-1)s_x^2}$ but σ_ε^2 is unknown, so we substitute the estimate s_E^2 and write

$$\hat{\mathrm{Var}}\left(\hat{\beta}\right) = \frac{s_E^2}{(n-1)s_x^2} \quad \text{and} \quad \text{s.e.}\left(\hat{\beta}\right) = \sqrt{\frac{s_E^2}{(n-1)s_x^2}}$$

Defn: The *standard error* of a statistics is an *estimate* of the *standard deviation of its sampling distribution*.

Question: What is the definition of a *sampling distribution*?

*** R simulation of sampling distributions ***

Confidence intervals: (review?)

$$\frac{\hat{\beta}-\beta}{\sqrt{\sigma_\varepsilon^2 / \sum (x_i-\bar{x})^2}} \sim ?$$

$$\frac{\hat{\beta}-\beta}{\sqrt{\text{s.e.}(\hat{\beta})}} \sim t_{(n-2)}$$

$$\Leftrightarrow P\left(-t_{n-2,\alpha/2} \le \frac{\hat{\beta}-\beta}{\sqrt{\text{se}(\hat{\beta})}} \le t_{n-2,\alpha/2} \right) = 1-\alpha \quad (\text{where } t_{n-2,\alpha/2} \text{ is the } (1-\alpha/2)\ \%\text{-ile})$$

$$\Leftrightarrow P\left(\hat{\beta} - t_{n-2,\alpha/2}\,\text{se}(\hat{\beta}) \le \beta \le \hat{\beta} + t_{n-2,\alpha/2}\,\text{se}(\hat{\beta}) \right)$$

Same type of calculation applies, of course, for $\hat{\alpha}$.

Hypothesis test: (review?)

$$H_0 : \beta = \beta_0 \quad \text{(most often considering } \beta_0 = 0\text{)}$$

$$t_0 = \frac{\hat{\beta} - \beta_0}{\text{se}(\hat{\beta})} \sim t_{(n-2)} \quad \text{under } H_0$$

$$P_{H_0}\left(\left| \frac{\hat{\beta} - \beta_0}{\text{se}(\hat{\beta})} \right| > t_{(n-2),\alpha/2} \right) = \alpha$$

$$\Rightarrow \text{ reject if } t_0 > t_{(n-2),\alpha/2}$$

Example: Consider the Davis weight-reported weight data. Test $H_0 : \beta = 1$

Question: How would you test the null hypothesis $H_0 : \beta = 1$ and $\alpha = 0$?

(there are two computational approaches that provide the same answer)

6.2 Multiple Regression

The mathematics becomes a little more interesting for multiple regression problems. (Sects 6.2, 9.3).

Let's first go back and revisit the least squares solution of chapter 5 in a little more detail.

$$RSS = \|\boldsymbol{\varepsilon}\|^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$
$$= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

Results for differenctiation of a matrix expression:

$$\frac{\partial(\mathbf{c}^T\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \begin{pmatrix} \partial(\mathbf{c}^T\boldsymbol{\beta})/\partial\beta_1 \\ \vdots \\ \partial(\mathbf{c}^T\boldsymbol{\beta})/\partial\beta_{k+1} \end{pmatrix} = \begin{pmatrix} c_1 \\ \vdots \\ c_{k+1} \end{pmatrix} = \mathbf{c}$$

$$\frac{\partial(\boldsymbol{\beta}^T\mathbf{C}\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = 2\mathbf{C}\boldsymbol{\beta}$$

which leads to

$$\frac{\partial \mathrm{RSS}}{\partial\boldsymbol{\beta}} = 2\mathbf{X}^T\mathbf{y} - 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = 0$$

or

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

This works if $\left(\mathbf{X}^T\mathbf{X}\right)$ has an inverse, which means that $\mathbf{X}$ must be have "full column rank". That is, the columns of $\mathbf{X}$ must be *linearly independent*: no column can be an exact linear combination of the other columns. We'll see shortly that there are consequences for columns which are nearly linearly dependent.

Next consider the basic properties in matrix notation:

$$\begin{aligned}\mathrm{E}(\hat{\boldsymbol{\beta}}) &= \left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\right) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{E}(\mathbf{y}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \\ \mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \mathrm{Var}(\mathbf{M}\mathbf{y}) = \mathbf{M}\,\mathrm{Var}(\mathbf{y})\,\mathbf{M}^T\end{aligned}$$

This is a matrix of variances and covariances. We are simplifying notation setting $\mathbf{M} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}$.

Because $\mathrm{Var}(\mathbf{y}) = \sigma_\varepsilon^2\mathbf{I}_n$, where $\mathbf{I}_n$ is the $n \times n$ identify matrix, the expression above simplifies to

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma_\varepsilon^2\mathbf{M}\,\mathbf{M}^T = \sigma_\varepsilon^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \sigma_\varepsilon^2(\mathbf{X}^T\mathbf{X})^{-1}$$

You should know and be comfortable with the derivation of this fundamental expression.

Under normality, $\hat{\boldsymbol{\beta}} \sim \boldsymbol{N}_{k+1}(\boldsymbol{\beta}, \sigma_\varepsilon^2(\mathbf{X}^T\mathbf{X})^{-1}$.

The *standard errors of the estimated regression coefficients* are the square roots of the diagonal elements of $s_E^2(\mathbf{X}^T\mathbf{X})^{-1}$.

There is a very useful expression for the $\mathrm{Var}(\hat{\beta}_j)$, the jth diagonal element of $\sigma_\varepsilon^2(\mathbf{X}^T\mathbf{X})^{-1}$ that is reported as eqn (6.2). It can be shown (but you will not be required to show) that

$$\mathrm{Var}(\hat{\beta}_j) = \frac{1}{\sqrt{1-R_j^2}} \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_{ij} - \overline{x}_i)^2}$$

$$= \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_{ij} - \hat{x}_{ij})^2}$$

where R_j^2 is the *squared multiple correlation coefficient* for regression of $\mathbf{x}_j$, the jth column of $\mathbf{X}$, on all the other columns of $\mathbf{X}$, and $\hat{x}_{ij}$ represents the fitted values from this auxiliary regression.

The term $\frac{1}{\sqrt{1-R_j^2}}$ is called the *variance-inflation factor*. If this factor is large, meaning that the jth column of $\mathbf{X}$ is well-predicted by (highly correlated with) the other columns of $\mathbf{X}$, the uncertainty in $\hat{\beta}_j$ will be large. The second line of the expression for $\mathrm{Var}(\hat{\beta}_j)$ tells us the same thing: if the jth column of $\mathbf{X}$ is well-predicted by the other columns of $\mathbf{X}$, the *conditional* variation in $\mathbf{x}_j$ given its prediction $\hat{\mathbf{x}}_j$ by the other columns, $\sum_{i=1}^n (x_{ij} - \hat{x}_{ij})^2$, will be small and, hence, $\mathrm{Var}(\hat{\beta}_j)$ will be large. *That is*, we will have a problem if the columns of $\mathbf{X}$ are nearly linearly dependent.

Tests of $H_0 : \beta_j = 0$ and confidence intervals for β_j are computed exactly as in simple linear regression, but using the standard errors of $\hat{\beta}_j$ presented above.

Table 9.1 Comparison between simple regression using scalars and multiple regression using matrices

	Simple Linear Regression	Multiple Linear Regression
Model	$Y = \alpha + \beta x + \varepsilon$	$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
Least-squares estimator	$\hat{\beta} = \frac{\sum x^* Y^*}{\sum (x^*)^2}$	$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{y}$
Sampling variance	$V(\hat{\beta}) = \frac{\sigma_\varepsilon^2}{\sum (x^*)^2} = \sigma_\varepsilon^2\left(\sum (x^*)^2\right)^{-1}$	$\mathrm{V}(\hat{\boldsymbol{\beta}}) = \sigma_\varepsilon^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$
Distribution	$\hat{\beta} \sim N\left(\beta, \sigma_\varepsilon^2\left(\sum (x^*)^2\right)^{-1}\right)$	$\hat{\boldsymbol{\beta}} \sim N_{k+1}\left(\underset{\sim}{\boldsymbol{\beta}}, \sigma_\varepsilon^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)$

Interpretation of the multiple regression model $Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

- In a *simple linear regression* (one explanatory variable), the coefficient β is called a ***marginal*** regression coefficient

- In a *multiple linear regression*, the coefficient β_j is called the ***partial*** regression coefficient—the "effect" on the response variable of a one unit change in the corresponding explanatory variable ***holding constant*** the values of all other explanatory variables. *[Does this always make sense?]*

- The least squares coefficient represents the average change in Y associated with a one unit change in X_j when all the other $X's$ are held constant.

R example: Prestige ~ Income + Education + . . .

The ANOVA for multiple regression. As shown in chapter 5

$$\text{TSS} = \text{RegSS} + \text{RSS}$$

where

$$\text{TSS} = \sum \left(Y_i - \bar{Y}\right)^2$$
$$\text{RegSS} = \sum \left(\hat{Y}_i - \bar{Y}\right)^2$$
$$\text{RSS} = \sum \left(Y_i - \hat{Y}\right)^2$$

The conventional table is:

Source	SS	Df	MS	F
Regression	RegSS	k	RegMS=RegSS/k	RegMS/RMS
Residual	RSS	n-k-1	RMS=RSS/(n-k-1)	
Total	TSS	n-1		

Under $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k$, $F = \dfrac{\text{RegMS}}{\text{RSS}} = \dfrac{\text{"MSReg"}}{\text{"MSE"}} \sim F_{k,n-k-1}$

Why? The F statistic is the ratio of two independent mean squares (they are independent because they derive from squared lengths of orthogonal vectors in a geometric representation) divided by their degrees of freedom:

$$F = \frac{\chi^2_k / k}{\chi^2_{n-k-1} / (n-k-1)}$$

In general, $\text{E(RegMS)} = \sigma^2_\varepsilon + f(\beta's)$, and under $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k$, $\text{E(RegMS)} = \sigma^2_\varepsilon$

We always have $\text{E(RMS)} = \sigma^2_\varepsilon$, so under H_0 we have the ratio of two independent chi-square random variables, both estimating the same thing, and this is the definition of a *central* F-distribution.

A little algebra shows that we can also express the F-statistic as

$$F = \frac{n-k-1}{k} \times \frac{R^2}{1-R^2}$$

It really helps to see this in matrix notation: $\mathbf{y} = \mathbf{X}\beta + \varepsilon,\ \hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$

In Chap 11 you'll see the notation $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ and the residuals are $(\mathbf{y} - \hat{\mathbf{y}}) = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}$ is commonly referred to as the "hat matrix". It is a *projection matrix*: $\mathbf{H}\mathbf{H} = \mathbf{H}$, and similarly $(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = (\mathbf{I} - \mathbf{H})$. And, $\mathbf{H}(\mathbf{I} - \mathbf{H}) = \mathbf{0}$, so these two projection matrices are *orthogonal*.

This is illustrated in Fig 10.6 for a regression with k=2 explanatory variables and the problem represented in mean-centered form. $\hat{\mathbf{y}}^* = \mathbf{H}\mathbf{y}^*$, where $\mathbf{H}$ is the projection matrix defined from the design matrix $\mathbf{X} = \left[\mathbf{x}_1^* \mathbf{x}_2^*\right]$.

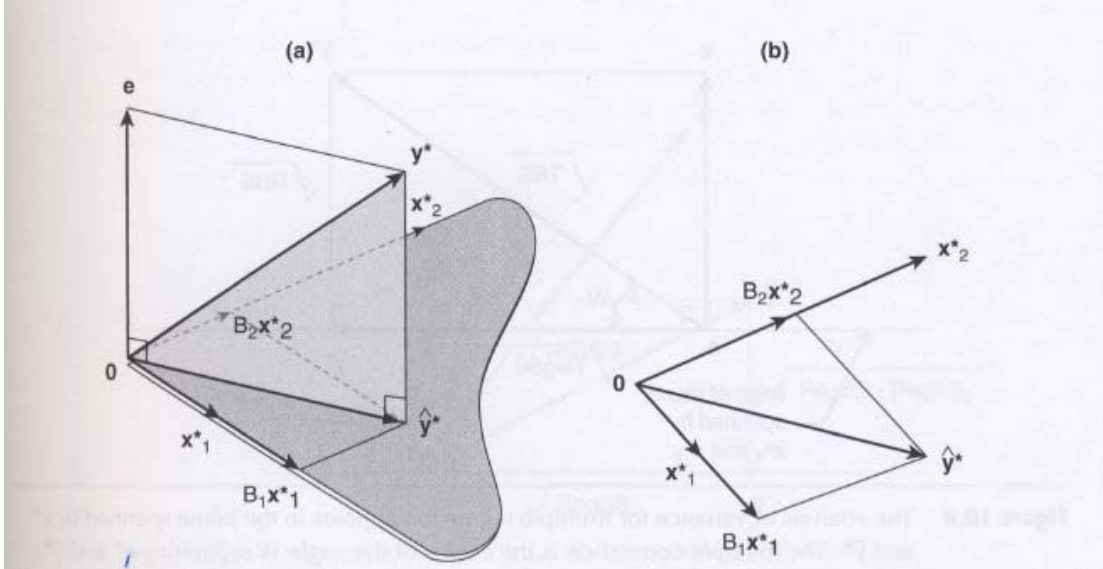


Figure 10.6 The vector geometry of least-squares fit in multiple regression, with the variables in mean-deviation form. The vectors $\mathbf{y}^*$, $\mathbf{x}_1^*$, and $\mathbf{x}_2^*$ span a three-dimensional subspace, shown in (a). The fitted-*Y* vector, $\widehat{\mathbf{y}}^*$, is the orthogonal projection of $\mathbf{y}^*$ onto the plane spanned by $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$. The $\{\mathbf{x}_1^*, \mathbf{x}_2^*\}$ plane is shown in (b).

(Dropping the notation for the centered variables)

$$\begin{aligned}\mathbf{y}^T\mathbf{y} &= \left(\hat{\mathbf{y}} + (\mathbf{y} - \hat{\mathbf{y}})\right)^T \left(\hat{\mathbf{y}} + (\mathbf{y} - \hat{\mathbf{y}})\right) \\ &= \left(\mathbf{H}\mathbf{y} + (\mathbf{I} - \mathbf{H})\mathbf{y}\right)^T \left(\mathbf{H}\mathbf{y} + (\mathbf{I} - \mathbf{H})\mathbf{y}\right) \\ &= \mathbf{y}^T\mathbf{H}\mathbf{y} + \mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{y} = \text{RegSS} + \text{RSS}\end{aligned}$$

Note that the Regression SS can also be written

$$\mathbf{y}^T\mathbf{H}\mathbf{y} = \mathbf{y}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y}$$

The ANOVA table can therefore be written

Source	SS	Df	MS
Regression	$\mathbf{y}^T\mathbf{H}\mathbf{y} = \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y}$	k+1	RegMS=RegSS/(k+1)
Residual	$\mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y}$	n-k	RMS=RSS/(n-k)
Total (“uncorrected”)	$\mathbf{y}^T\mathbf{y}$	n-1	

If we write everything in mean-centered form, the same expressions hold (with *’s on all the vectors and matrices to denote mean-centering) with k and (n-k-1) df. The “Total” sum of squares is then “corrected for the mean”, $\mathbf{y}^{*T}\mathbf{y}^* = \sum (y_i - \bar{y})^2$.

The *omnibus* F-test derived from this ANOVA is a test of the significance of the multiple regression, rejected if there is evidence that at least one of the $\beta's$ is significantly different from zero.

A modest generalization of this provides a basis for *testing a subset of the slopes*. In the general framework we define

RSS_1 and RegSS_1 the regression sum of squares for the *full model*

RSS_0 and RegSS_0 the residual and regression sums of squares for the *reduced model*

and

$$F = \frac{\left(\mathrm{RegSS}_1 - \mathrm{RegSS}_0\right)/q}{\mathrm{RSS}_1/(n-k-1)} = \frac{\left(\mathrm{RSS}_0 - \mathrm{RSS}_1\right)/q}{\mathrm{RSS}_1/(n-k-1)}$$

In matrix notation,

$$\mathbf{y} = \boldsymbol{\alpha} + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\beta}_1 = \left(\beta_1,\ldots,\beta_q\right)^T$, $\boldsymbol{\beta}_2 = \left(\beta_{q+1},\ldots,\beta_k\right)^T$ and $\mathbf{X} = \left[\mathbf{X}_1\ \mathbf{X}_2\right]$ with $\mathbf{X}_1$ and $\mathbf{X}_2$ having q and $(k-q)$ columns, respectively.

Let

$\mathrm{RegSS}_0 = \mathrm{SS}(\hat{\boldsymbol{\beta}}_2) = \hat{\boldsymbol{\beta}}_2\mathbf{X}_2^T\mathbf{y}$, the regression sum of squares for the *reduced model*.

$\mathrm{RegSS}_1 = \mathrm{SS}(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) = \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y}$, the regression sum of squares for the *full model*.

Under $H_0 : \beta_1 = 0$, $F = \dfrac{\left(\mathrm{SS}(\hat{\beta}_1, \hat{\beta}_2) - \mathrm{SS}(\hat{\beta}_2)\right)/q}{\mathrm{RSS}/(n-k-1)} \sim F_{q,(n-k-1)}$

***If** $\mathbf{X}_1$ and $\mathbf{X}_2$ are orthogonal, then

$$\mathrm{SS}(\hat{\beta}_1, \hat{\beta}_2) = \hat{\beta}_1 \mathbf{X}_1^T \mathbf{y} + \hat{\beta}_2 \mathbf{X}_2^T \mathbf{y}$$

Source	SS	Df	MS
$\mathbf{X}_2 : SS(\hat{\beta}_2)$	$\hat{\beta}_2{}^T \mathbf{X}_2{}^T \mathbf{y}$	$k-q$	RegMS_2
$\mathbf{X}_1 \vert \mathbf{X}_2 : SS(\hat{\beta}_1 \vert \hat{\beta}_2)$	$\hat{\beta}^T \mathbf{X}^T \mathbf{y} - \hat{\beta}_2{}^T \mathbf{X}_2{}^T \mathbf{y}$	q	$\mathrm{RegMS}_{1\vert 2}$
Residual	$\mathbf{y}^T (\mathbf{I} - \mathbf{H}) \mathbf{y} = \mathbf{y}^T \mathbf{y} - \hat{\beta}^T \mathbf{X}^T \mathbf{y}$	n-k	RMS=RSS/(n-k)
Total	$\mathbf{y}^{*T} \mathbf{y}^*$	n-1	

* A special case for testing $q = 1$ coefficient, $H_0 : \beta_1 = 0$. F-test or t-test?

$F_{1,(n-k-1)} = t^2_{(n-k-1)}$. Because $F = \dfrac{n-k-1}{k} \times \dfrac{R^2}{1-R^2}$, $t = \sqrt{n-2}\,\dfrac{r}{\sqrt{1-r^2}}$.

Return to R code and anova() for multiple regression.

6.3 Empirical vs Structural Relations

Empirical: Descriptive relationship among variables

Structural: Descriptions from which we intend to infer *causation*, a model of how response scores are actually determined.

(a) $Y = \alpha' + \beta_1' X_1 + \varepsilon'$

(b) $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + +\varepsilon$

If the relationship between Y and X_1 is well described by the simple linear relationship (a), the fact that $\beta_1' \neq \beta_1$, so that $\hat{\beta}_1'$ may be considered a *biased estimate* of β_1, is not necessarily an issue. However, if we really believed that the data were generated by model (b), this bias would be important.

If (b) is true, but we fit (a), then $\varepsilon' = \left(\beta_2 X_2 + \varepsilon\right)$, so that X_1 and ε' are correlated if X_1 and X_2 are correlated.

Correlation between the error term and the explanatory variable is a problem leading to bias of $\hat{\beta}_1'$ with respect to the value of β_1 in (b).

On pp. 111-112 Fox derives an expression for β_1 in model (b)

Take expected values of (b), giving

- $\mu_y = \alpha + \beta_1\mu_1 + \beta_2\mu_2 + 0$

and subtract this from (b), giving

- $\left(Y - \mu_y\right) = \beta_1\left(X_1 - \mu_1\right) + \beta_2\left(X_2 - \mu_2\right) + \varepsilon$

Multiply by $\left(X_1 - \mu_1\right)$, giving

- $\left(X_1 - \mu_1\right)\left(Y - \mu_y\right) = \beta_1\left(X_1 - \mu_1\right)^2 + \beta_2\left(X_1 - \mu_1\right)\left(X_2 - \mu_1\right) + \left(X_1 - \mu_1\right)\varepsilon$

Take expected value of both sides of this equation:

- $\sigma_{1Y} = \beta_1\sigma_1^2 + \beta_2\sigma_{12}$

Thus, solving for β_1,

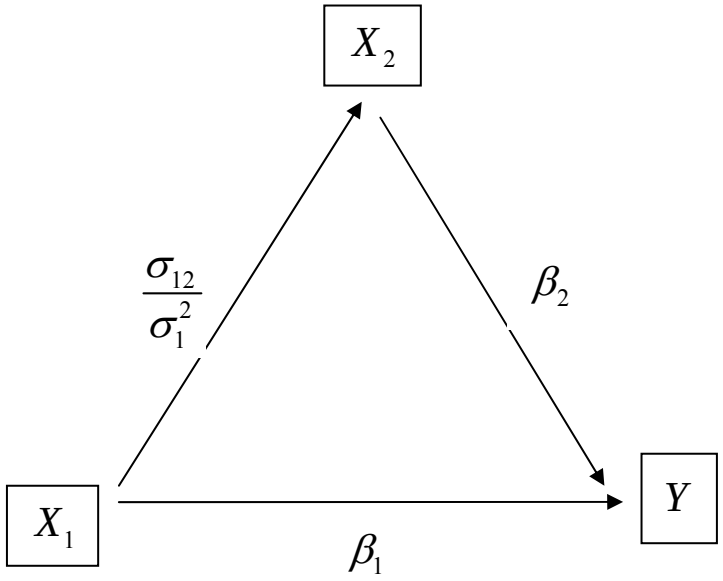
- $\beta_1 = \frac{\sigma_{1Y}}{\sigma_1^2} - \beta_2\frac{\sigma_{12}}{\sigma_1^2}$

The least squares estimate from model (a) is $\hat{\beta}_1' = \frac{s_{1Y}}{s_1^2}$, so that $\hat{\beta}_1'$ is an estimate of $\frac{\sigma_{1Y}}{\sigma_1^2} = \beta_1'$, meaning it is <u>biased</u> as an estimator of the coefficient for model (b), $\beta_1 = \frac{\sigma_{1Y}}{\sigma_1^2} - \beta_2\frac{\sigma_{12}}{\sigma_1^2}$. I.e., $\mathrm{E}\left(\hat{\beta}_1'\right) = \beta_1 + \beta_2\frac{\sigma_{12}}{\sigma_1^2}$.

We have bias if

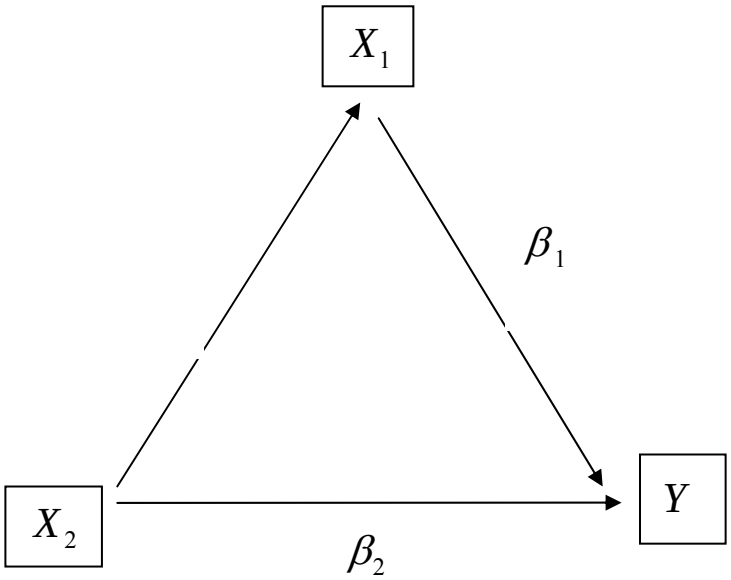
1. X_2 is "relevant" (i.e., $\beta_2 \neq 0$), and
2. X_1 and X_2 are correlated.

One final subtlety of interpretation: the "bias" $\beta_2 \frac{\sigma_{12}}{\sigma_1^2}$ depends on the nature of the causal relationship between X_1 and X_2. In Fig 6.2(a) $\beta_2 \frac{\sigma_{12}}{\sigma_1^2}$ is seen to be the *indirect effect* of X_1 on Y through X_2.



$\frac{\sigma_{12}}{\sigma_1^2}$ is the population slope for the regression of X_2 on X_1. X_2 is an *intervening variable* on a causal path through X_2.

By contrast, in Fig 6.2(b), X_2 is a *common prior cause* of both X_1 and Y.



In this case the bias is considered a *spurious* (non-causal) component of the association between X_1 and X_2.

And in this case it is critical to include the variable X_2 in the analysis— "*to control for* X_2" in examining the association between X_1 and X_2.

X_2 is sometimes described as a *lurking variable*. These are what make causal inference from observational studies so difficult.

Omission of X_2 is not necessarily (always) an issue in the case of Fig 6.2(a), depending on whether it is important to understand the direct and indirect effects of X_1.

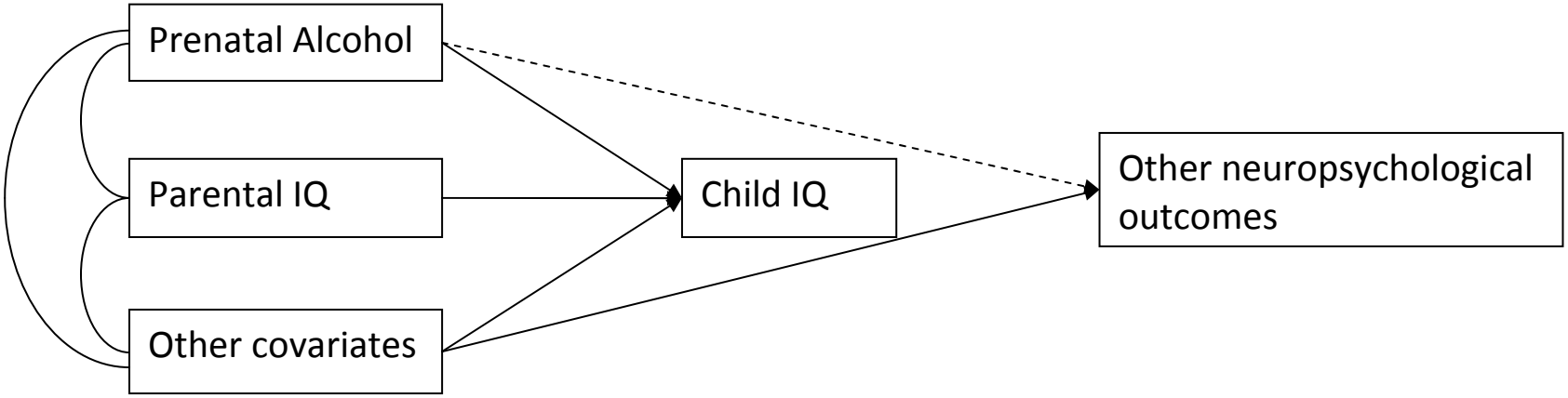
So, you have to think hard about your problem if the scientific question(s) target the effect of a particular explanatory variable "X_1".

Examples:

- X_1 is a measure of Chinese investment in the U.S. Y is a measure of U.S. stock prices. "X_2"?
- X_1 is an indicator distinguishing two ethnic groups. Y is a measure of public health outcome. "X_2"?

See Fox Exercise 6.8 in addition to Exercise 6.9 assigned for Homework #4.

Another example from my applied work: The effects of Prenatal Alcohol exposure on offspring IQ.



We can't say that Parental IQ causes Maternal drinking (prenatal alcohol exposure), but it is correlated with prenatal alcohol (in a manner that depends on how you measure prenatal alcohol), so it contributes to the observed association of Prenatal Alcohol with Child IQ.

$\Rightarrow$ you must have Parental IQ in the model or else you will have a biased estimate of the effect of Prenatal Alcohol on Child IQ.

What do you do if you haven't measured Parental IQ?

6.4 Measurement Error in Explanatory Variables

The basic multiple regression models we work with assume that the explanatory variables are known without error. Sometimes this is a reasonable assumption and sometimes not. Suppose you believe in a multiple regression model as expressed in eqn (6.9),

$$Y = \beta_1 \tau + \beta_2 X_2 + \varepsilon$$

where X_2 is measured without error, but you cannot directly observe τ. Instead you have a measurement of τ (a "fallible indicator")

$$X_1 = \tau + \delta$$

where δ represents measurement error. And we are particulary interested in the coefficient β_1 (not just whether our multiple regression does a good job of predicting Y).

I will not go through the algebra presented in this section as we will not address the approaches to deal with the subject of measurement error. Suffice to say that, when you compute the regression of Y on X_1 and X_2, the measurement error in X_1

- "*attenuates*" the coefficient estimate $\hat{\beta}_1$ — makes it systematically biased (smaller in absolute value) with respect to the coefficient β_1 in the true model.
- can bias the coefficient estimate $\hat{\beta}_2$ in either a positive or negative direction (toward or away from 0) depending on sign of covariance between X_1 and X_2.

Explanation of effect of measurement error in the case of simple linear regression (not from Fox).

$$Y_i = \alpha + \beta_1 \tau_i + \varepsilon_i, \quad X_i = \tau_i + \delta_i$$
$$= \alpha + \beta_1 X_i + (\varepsilon_i + \beta_1 \delta_i) = \alpha + \beta_1 \tau_i + \varepsilon_i^*$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2), \quad \delta_i \sim N(0, \sigma_\delta^2)$.

Define $\sigma_\tau^2 = \sum (\tau_i - \bar{\tau})^2 / n$, the variance in the "true" predictor, and assume assume variability in τ_i is uncorrelated with the measurement error δ_i.

It is an easy exercise to show that

$$\mathrm{E}(\hat{\beta}_1) = \mathrm{B}_1 \Big/ \left(1 + \sigma_\delta^2 / \sigma_\tau^2\right) < \beta_1$$

The problem occurs because X_i and ε_i^* are usually not independent for this model

$$\mathrm{Cov}(X_i, \varepsilon_i^*) = \mathrm{Cov}(\tau_i + \delta_i, \varepsilon_i - \beta_1 \delta_i) = -\beta_1(\sigma_{\tau\delta} + \sigma_\delta^2)$$

Can generally ignore measurement error if

1) $\sigma_\delta^2 / \sigma_\tau^2$ is small --- measurement error small relative to true variance in τ's
2) X_i are fixed and predetermined by design ($\tau_i = X_i - \delta_i, \quad \sigma_{\tau\delta} + \sigma_\delta^2 = 0$)
3) postulated model is $Y_i = \alpha + \beta_1 X_i + \varepsilon_i$

You can probably imagine how measurement error must be a consideration in may social science models where explanatory variables are *concepts* measured with error.

What is "education" in the occupational prestige dataset? Is it a concept measured with error? See Exercise 6.13.

Example of practical importance of measurement error in our current research on effects of air pollution on public health.

- Large EPA-funded study of effects of long-term exposure to "fine particulate matter" ($PM_{2.5}$) and other air pollutants, especially traffic-related pollutants) on cardiovascular disease in older people.
- We want to know the effect of "true exposure" to $PM_{2.5}$ on measures of coronary artery calcium, for example, as well as cardiovascular "events". This has great implications for public health and regulation of air pollution.
- $PM_{2.5}$ concentration is measured over time at a fixed number of monitoring sites and these data are used to estimate concentrations at the locations where the subjects live. There is considerable error in the estimation of true concentration at the locations where people live (let alone "true exposure", which involves lots of other factors).
- It is a current research activity to determine good ways to deal with measurement error in this problem.