Chapter 6 Multiple Regression

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:

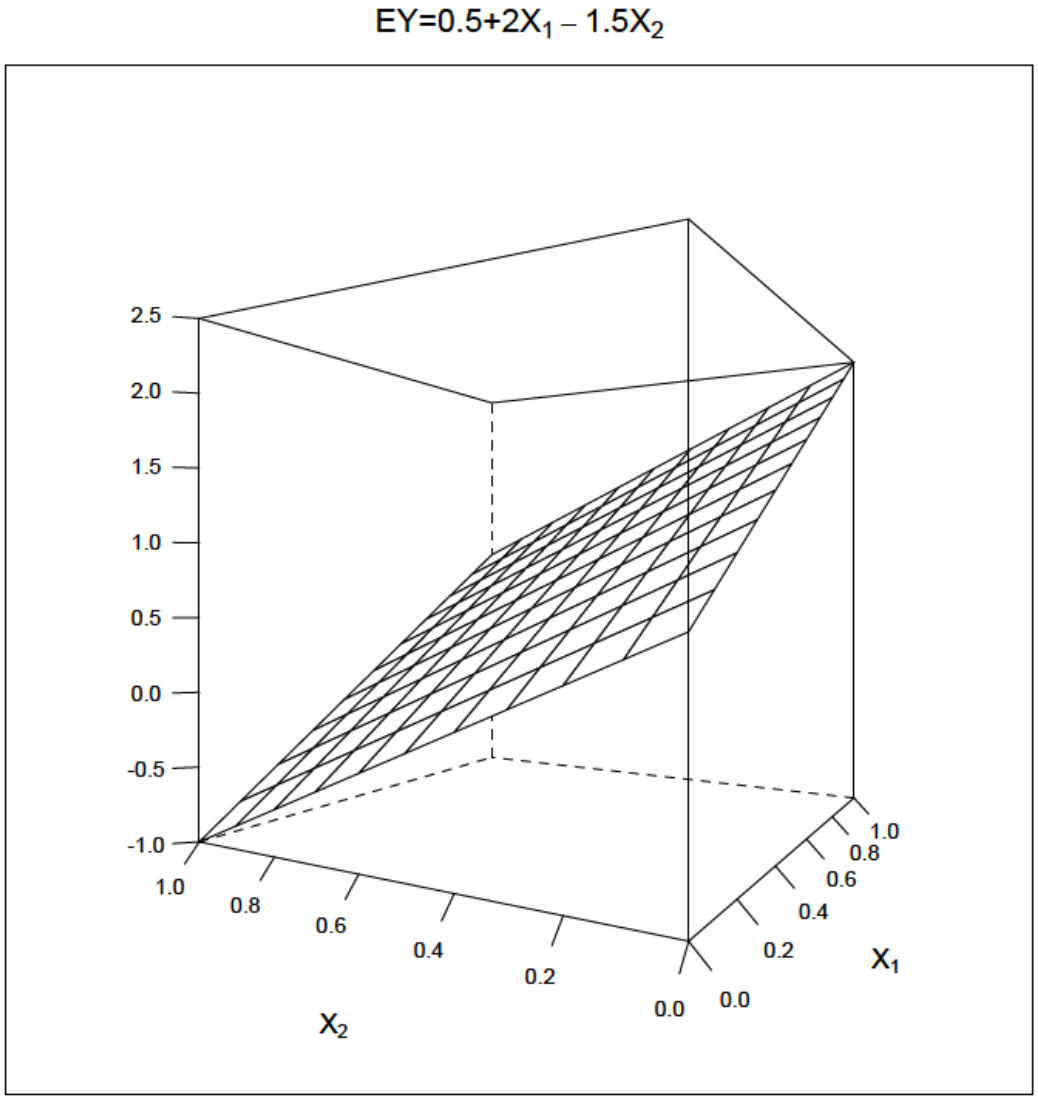
$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i,$$

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 $\mathbb{R}^3$.

Multiple regression models



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:

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

2009 Water Quality Data Set

Water samples from tributaries of the Congaree River. E Coli is to be replaced by another measure of bacterial water quality.

- $Y =$ log E Coli count (colonies/ml H_2O)
- $x_1 =$ log Fecal Coliform count (colonies/ml H_2O)
- $x_2 =$ log Enterococci count (colonies/ml H_2O)

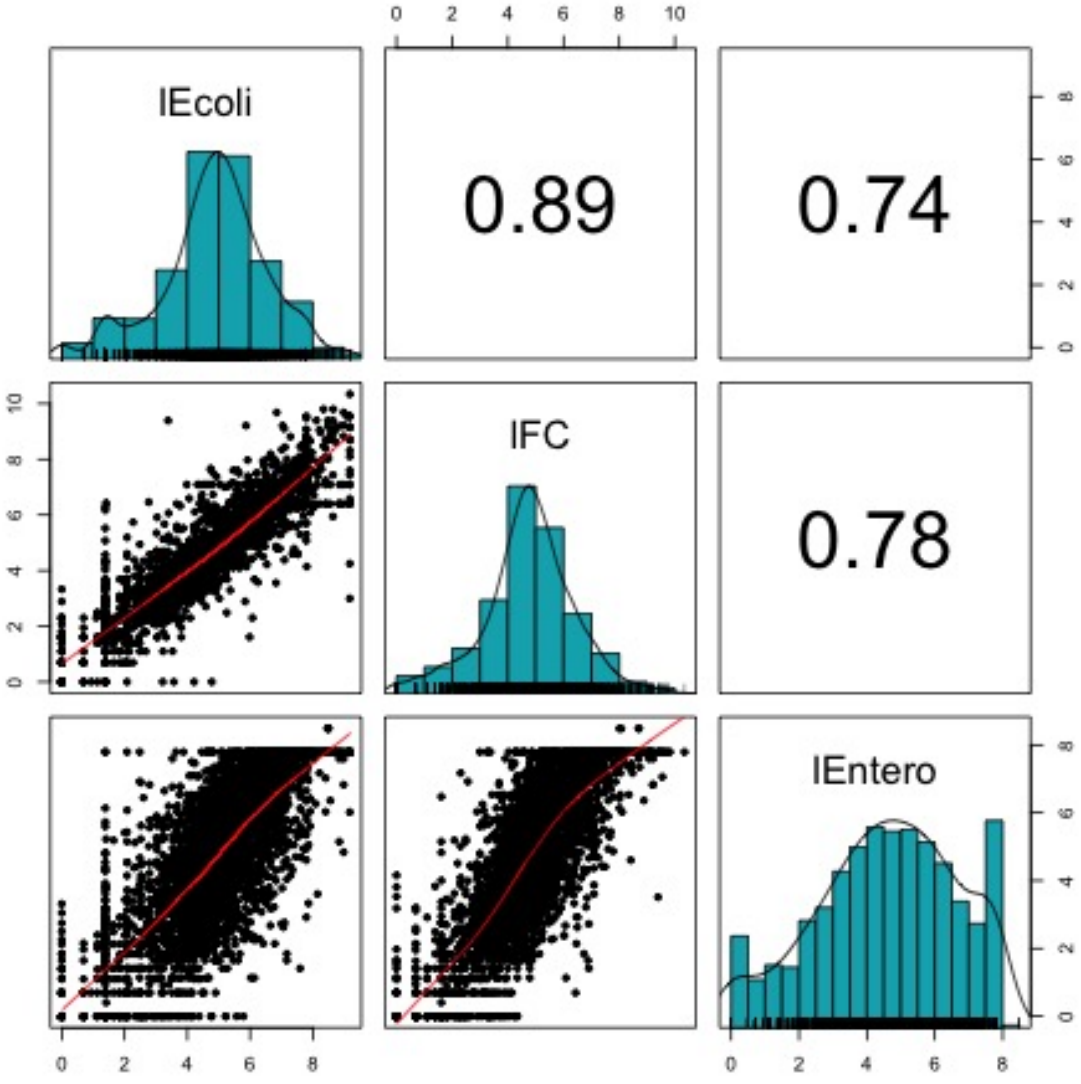
Assume the linear model is appropriate. One way to check marginal relationships is through a scatterplot matrix. However, these are not infallible.

β_2 is the change in the mean log response for a 1 log-colony increase in Enterococcus, holding "Fecal Coliform log count" constant.

R code

```
ecoli<-read.csv("EColi.csv", header=T, stringsAsFactors=F)
str(ecoli)
lEcoli<-log(ecoli$Ecoli)
lFC<-log(ecoli$FecalColi)
lEntero<-log(ecoli$Enterococci)
dat<-data.frame(lEcoli, lFC, lEntero)
library(psych)
pairs.panels(dat,
             method = "pearson", # correlation method
             hist.col = "#00AFBB",
             density = TRUE,  # show density plots
             ellipses = F # show correlation ellipses
)
```

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 $x_2 = x_1^2$ places this 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 $Y^* = \log(Y)$ to obtain a 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?

General linear model in matrix terms

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

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

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

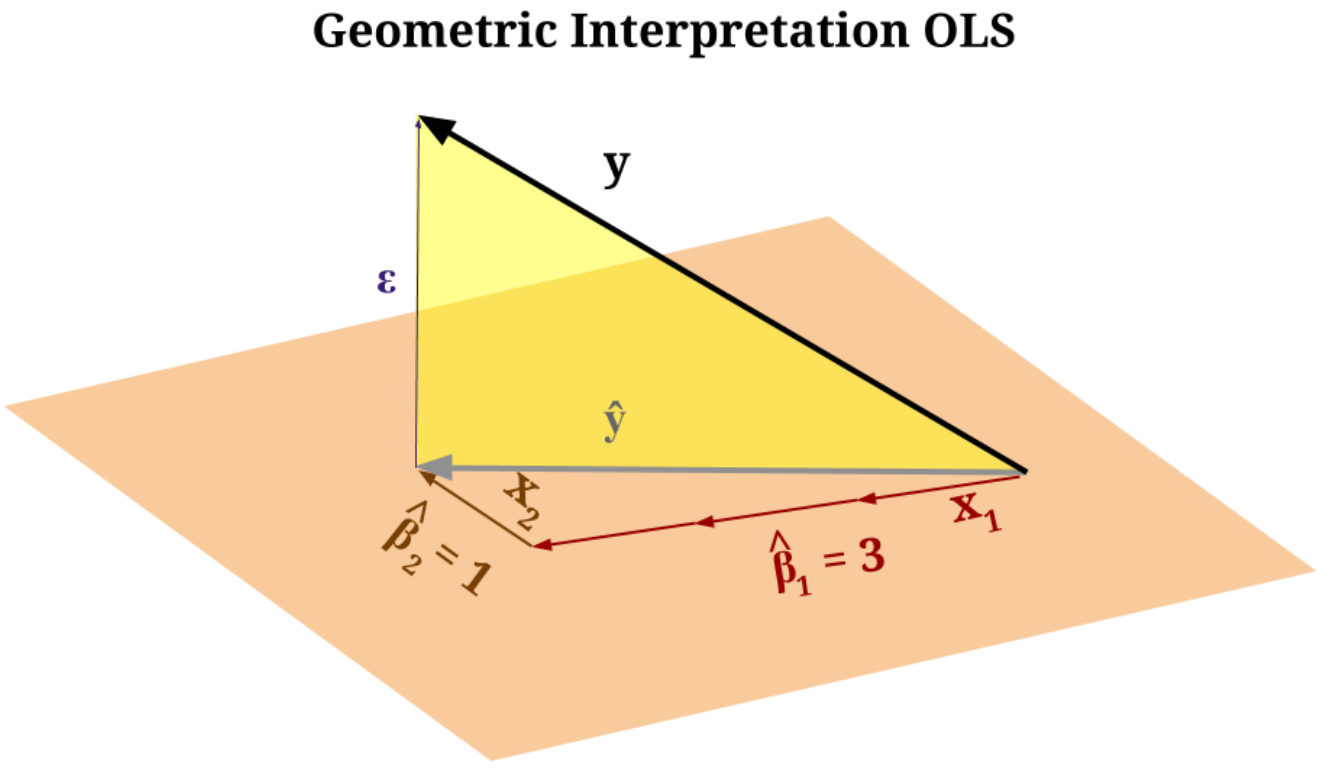
The *fitted values* are in the vector

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

The *residuals* are in the vector

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

Geometry



$$Cov(\epsilon, \hat{\mathbf{Y}}) = Cov[(\mathbf{I} - \mathbf{H})\mathbf{Y}, \mathbf{HY}] = 0$$

Hat matrix

$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ 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 **Congaree water quality data**. From R,

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 0.52045 \\ 0.83149 \\ 0.06155 \end{bmatrix},$$

so the fitted regression line is

$$\hat{Y} = 0.52045 + 0.83149x_1 + 0.06155x_2.$$

Interpretation

- **Interpretation of** b_1: We *estimate* that for each one-unit increase in log(Fecal coliform), mean log E Coli increases by 0.83149 log counts when log(Enterococci) is held constant.
- **Interpretation of** $\exp\{b_1\}$: We *estimate* that for a one-colony increase in Fecal coliform, mean E Coli increases by $\exp\{0.83149\} = 2.30$ colonies when log(Enterococci) is held constant.
- **Interpretation of** b_2: We *estimate* that for each one-unit increase in log(Enterococci), mean log E Coli increases by 0.06155 log counts when log(Fecal coliform is held constant. **Note:** This is nonsensical.

Analysis of variance (ANOVA) table

Restated: The variation in the data (SSTO) can be divided into two parts: the part explained by the model (SSR), and the slop that's left over, i.e. unexplained variability (SSE).
Associated with each sum of squares are their degrees of freedom (df) and mean squares, forming a nice table:

Source	SS	df	MS	$E(MS)$
Regression	SSR= $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	1	$\frac{SSR}{1}$	$\sigma^2 + \beta_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2$
Error	SSE= $\sum_{i=1}^{n}(Y_i - \hat{Y})^2$	$n-2$	$\frac{SSE}{n-2}$	σ^2
Total	SSTO= $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$	$n-1$		

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

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

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 $x_1, \ldots, x_k$ 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 $x_1, \ldots, x_k$ and Y. R 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.

Congaree water quality, ANOVA table, R^2, & R_a^2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	405.94	202.97	349.23	<.0001
Error	251	129.23	0.51		
Corrected Total	253	535.17			

Root MSE	0.7175	R-Square	0.7585
Dependent Mean	4.687	Adj R-Sq	0.7566
Coeff Var	15.3084		

We reject $H_0 : \beta_1 = \beta_2 = 0$ at any reasonable significance level α. About 76% 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 (Least Absolute Shrinkage and Selection Operator), LARS (Least-Angle Regression), etc. These are used when there's *lots* of predictors, e.g. $p = 500$ or $p = 20,000$.

Multivariate normal

The *multivariate normal* density is given by

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

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

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

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

For the general linear model,

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

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,

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

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

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

Back to the general linear model

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

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

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

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

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

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

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

Mean and variance of $\mathbf{b}$ (p. 227)

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

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

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

Table of regression effects (p. 228)

From the previous slide, the jth estimated coefficient β_j,

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

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 $\text{se}(b_j) = \sqrt{MSE c_{jj}}$ yielding

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

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

Congaree water quality output

- The R summary gives us $F^* = MSR/MSE = 394.23$ with associated p-value < 0.0001. We strongly reject (at any reasonable α) $H_0 : \beta_1 = \beta_2 = 0$.
- 95% CI's are $(0.73202, 0.93096)$ for β_1 and $(-0.00705, 0.13016)$ for β_2.
- For example, we are 95% confident that mean log E Coli count increases by 0.73202 to 0.93096 for every one log count increase in fecal coliform, holding enterococci constant.
- For $H_0 : \beta_1 = 0$ we get $p < 0.0001$; for $H_0 : \beta_2 = 0$ we get $p = 0.08$.