APPENDIX A
Review of Linear Regression and Multiple Linear Regression Analysis

In this appendix we review the concepts of linear regression and multiple linear regression analysis.

A.1 INTRODUCTION

Let's begin this review by discussing the **linear regression model**, based on a single covariate, an independent "predictor" variable x, and a response, a dependent variable y, with a linear relationship given by

$$y_i = \beta_0 + \beta_1 \cdot x_i + e_i,$$

where $e_i \sim$ iid $N(0, \sigma)$ are independent and identically distributed error (stochastic) elements or "residuals," normally distributed with mean 0 and standard deviation σ. Assume the x to be fixed and measured without error. There are two parameters, β_0, β_1, along with σ, that can be estimated with estimates $b_0 = \hat{\beta}_0$ and $b_1 = \hat{\beta}_1$ along with $s = \hat{\sigma}$ in this linear regression model. The assumptions for the linear regression model are as follows for the conditional population $y|x = \{y|(x, y)$ is in the population$\}$:

1. $\mu_{y|x}=\beta_0+\beta_1 \cdot x$ (linearity of the mean).
2. $\sigma_{y|x}=\sigma$ is constant (homoscedasticity).
3. $y|_x$ is normally distributed: $y|_x \sim N(\mu_{y|x}, \sigma)$ (normality).
4. $y|_x$ are randomly sampled and independent, with x either randomly sampled or fixed, and measured without error.

The data consist of ordered pairs of x and y sample measurements (x_i, y_i), $i = 1,2, \ldots, n$, from the larger population.

Contemporary Bayesian and Frequentist Statistical Research Methods for Natural Resource Scientists. By Howard B. Stauffer

We can begin the analysis by examining the sample dataset graphically, looking at an (x, y) scatterplot for the linearity of the data. The **Pearson sample correlation statistic**

$$r = \hat{\rho} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}},$$

which estimates the **Pearson correlation parameter** in the case of a population of finite size N

$$\rho = \frac{\sum_{i=1}^{N} (x_i - \mu_x) \cdot (y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N} (x_i - \mu_x)^2} \cdot \sqrt{\sum_{i=1}^{N} (y_i - \mu_y)^2}},$$

should first be examined for statistical significance before a linear regression modeling analysis is conducted. Correlation, varying between -1 and $+1$, "measures" whether there is a significant linear relationship between the variables x and y. Values near $+1$ or -1 suggest a linear relationship with a positive or negative slope, respectively, whereas values near 0 are indicative of no linear relationship. The analyst should generally proceed with a linear regression modeling analysis only if there is significant nonzero correlation. The correlation can be tested for significance

$$H_0 \colon \rho = 0$$

by using the test statistic

$$t_s = r \cdot \sqrt{\frac{n-2}{1-r^2}}.$$

This test statistic is t_{n-2}-distributed, if the null hypothesis is true.

The linear regression model with a single covariate x can be generalized to the **multiple linear regression model** with $p \geq 1$ "independent" "predictor" covariates $\{x_1, x_2, \ldots, x_p\}$ and dependent response y

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \cdots + \beta_p \cdot x_{ip} + e_i,$$

where $e_i \sim$ iid $N(0, \sigma)$ are independent and identically distributed error (stochastic) elements or "residuals," normally distributed with mean 0 and standard deviation σ.

There are $k = p+1$ parameters, $\beta_0, \beta_1, \ldots,$ and β_p, along with σ, which can be estimated with estimates $b_0 = \hat{\beta}_0$, $b_1 = \hat{\beta}_1, \ldots,$ and $\hat{\beta}_p$, along with the residual standard error $s_{y|x} = \hat{\sigma}$, in this multiple linear regression model. All the linear regression results described in this appendix will also generalize to the multiple linear regression model (Seber 1977, Draper and Smith 1981, Manly 1994, Hocking 1996, Ryan 1997, Cook and Weisberg 1999).

A.2 LEAST-SQUARES FIT: THE LINEAR REGRESSION MODEL

For linear regression analysis, a least-squares (LS) method is used to fit the linear regression model to the sample data, minimizing the **goodness-of-fit profile** and providing unbiased, minimum variance estimators for the parameters β_0 and β_1. In Section 2.2, we examined other ways of fitting models to data: maximum-likelihood (ML) estimators that maximize the likelihood profile and Bayesian methods. In this chapter, with linear regression modeling, we focus on LS fit.

The **least-squares (LS) fit** of the linear regression model estimates the parameters β_0 and β_1 by minimizing the sum of squared residuals, the goodness-of-fit profile

$$\text{GOF} = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 \cdot x_i))^2 = \sum_{i=1}^{n} e_i^2$$

(Hilborn and Mangel 1997). Taking partial derivatives of the GOF profile with respect to the unknown parameters β_0 and β_1 and setting these partial derivatives equal to 0, the solution for the **parameter estimators** is obtained with the normal equations

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (x_i - \bar{x})} = \frac{\text{SP}_{xy}}{\text{SS}_x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}.$$

The second equation can also be obtained from the fact that the mean average or centroid $(\bar{x}, \bar{y})$ of the x and y sample values falls on the regression line. Estimators for the **standard errors** of $\hat{\beta}_1$ and $\hat{\beta}_0$ are given by

$$\text{se}_{\hat{\beta}_1} = \hat{\sigma} \cdot \sqrt{\frac{1}{(n-1) \cdot s_x^2}},$$

$$\text{se}_{\hat{\beta}_0} = \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1) \cdot s_x^2}},$$

where the estimator for the standard deviation of the residual error, the **residual standard error**, $\hat{\sigma} = s_{y|x}$ is

$$\hat{\sigma} = \sqrt{\frac{\text{GOF}}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 \cdot x_i))^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}} = s_{y|x.}$$

The estimators for β_0 and β_1 are **BLUE (the best linear unbiased estimators)**; that is, they are unbiased estimators of minimum variance.

Similarly, for multiple linear regression analysis, a LS method is used to fit the multiple linear regression model to sample data, minimizing the **goodness-of-fit profile** and providing unbiased, minimum variance estimators for the parameters, $\beta_0, \beta_1, \ldots, \beta_p$.

A.3 LINEAR REGRESSION AND MULTIPLE LINEAR REGRESSION STATISTICS

In this section we discuss eight leading statistics that form the basis for evaluating linear and multiple linear regression models:

1. The coefficient estimates and their significance using either confidence intervals or *t* tests
2. The coefficient of determination R^2
3. The residual standard error $s_{y|x}$
4. The ANOVA *F* test
5. The adjusted R^2
6. The Mallows C_p
7. The Akaike information criterion (AIC) and the corrected Akaike information criterion (AIC_c)
8. The Bayesian information criterion (BIC)

A.3.1 Estimates of Coefficients and Their Significance: Confidence Intervals and *t* Tests

The estimates for slope ($\hat{\beta}_1$) and *y* intercept ($\hat{\beta}_0$) for the linear regression model can be tested for statistical significance at the $P = 1 - \alpha$ level of confidence with the null

hypotheses respectively

$$H_0: \beta_0 = 0$$
$$H_0: \beta_1 = 0$$

in two equivalent ways (where the degrees of freedom df are given by $\mathrm{df} = n - 2$):

1. Determine whether 0 is outside the confidence intervals:

$$\hat{\beta}_0 \pm t_{\mathrm{df},1-\alpha/2} \cdot \mathrm{se}_{\hat{\beta}_0}$$
$$\hat{\beta}_1 \pm t_{\mathrm{df},1-\alpha/2} \cdot \mathrm{se}_{\hat{\beta}_1}$$

2. Examine the significance of the t statistics:

$$t_s = \frac{\hat{\beta}_0}{\mathrm{se}_{\hat{\beta}_0}}$$
$$= \frac{\hat{\beta}_1}{\mathrm{se}_{\hat{\beta}_1}}$$

Reject H_0 if and only if the confidence interval does not contain 0. This occurs if and only if the test statistic falls within the rejection region, or equivalently if the p value is less than or equal to the type I error α. Otherwise, do not reject H_0.

It is important to determine whether the estimates for β_0 and β_1 are statistical significant. If the estimate for β_1 is not statistically significant, then β_1 is statistically equivalent to 0 and the linear term for x should be eliminated, reducing the model to the **null model**:

$$y_i = \beta_0 + e_i.$$

If the estimate for β_0 is not statistically significant, then β_0 is statistically equivalent to 0 and the constant term should be eliminated from the model and the **ratio model** used instead if β_1 is statistically significant:

$$y_i = \beta_1 \cdot x_i + e_i.$$

A.3.2 The Coefficient of Determination R^2

The **coefficient of determination** R^2 is another important statistic used to evaluate the linear model with the LS fit. It varies between 0 and 1, with a higher R^2 value indicating a linear relationship. For the linear regression with one independent

variable, $R = r$, the Pearson correlation coefficient, whereas with multiple linear regression with more than one independent covariate, R generalizes the Pearson correlation coefficient to more than one dimension. The R^2 index "measures" the proportion of the variation of y "captured" by the regression model

$$R^2 = \frac{\mathrm{SS}_{\mathrm{regression}}}{\mathrm{SS}_{\mathrm{total}}} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$
$$= 1 - \frac{\mathrm{SS}_{\mathrm{error}}}{\mathrm{SS}_{\mathrm{total}}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

where the three components of variation are given by

$$\mathrm{SS}_{\mathrm{regression}} = \sum_i (\hat{y}_i - \bar{y})^2,$$
$$\mathrm{SS}_{\mathrm{error}} = \sum_i (y_i - \hat{y}_i)^2,$$
$$\mathrm{SS}_{\mathrm{total}} = SS_y = \sum_i (y_i - \bar{y})^2,$$

with

$$\mathrm{SS}_{\mathrm{total}} = \mathrm{SS}_y = \mathrm{SS}_{\mathrm{regression}} + \mathrm{SS}_{\mathrm{error}}$$

and $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$.

Figure A.1 displays scatterplots of (x,y) samples from populations satisfying the assumptions of linear models, with four prototype cases of R^2 values exhibited with extremes of slope and error. The highest R^2 values occur with datasets having maximal slope and minimal error. The lowest R^2 values occur with datasets having minimal slope and maximal error. Moderate R^2 values occur with datasets having moderate slope and moderate error. The R^2 index effectively "measures" both steepness and slope.

However, R^2 alone does not provide a definitive indicator of fit, despite common misperception. Datasets may have a high R^2, with a large amount of steepness and small amount of error, yet be curvilinear, rather than linear, in shape. Additionally, most natural resource datasets will contain an appreciable amount of stochastic error that cannot readily be explained by identifiable and measurable independent covariates, so there is a limit to the amount of error that a model should minimize. Yet a linear regression model may still prove useful and provide a reasonable fit, with a significant amount of error estimated by $s_{y|x}$. A natural resource relationship may be gradual, with moderate slope, and with a moderate amount of R^2 value, say, between 20% and 60%, yet still be statistically and biologically significant.

Models with high R^2 values may overfit sample datasets and not serve well as predictive models (Burnham and Anderson 1998). As the number of independent covariates in a model increases, R^2 values also increase, since the model will have

Figure A.1. Scatterplot of (x, y) sample datasets from populations satisfying the assumptions of linear regression, with extremes of slope and error, and their effects on the estimated coefficients of determination R^2. (**a**) High slope, low error: `x <- runif(80,0,100), y1 <- 30+5.0*x+rnorm(80,0,20)`, $R^2 = 0.992$. (**b**) Low slope, low error: `x <- runif(80,0,100), y2 <- 30+1.2*x+rnorm(80,0,20)`, $R^2 = 0.984$. (**c**) high slope, high error: `x <- runif(80,0,100), y3 <- 30+5.0*x+rnorm(80,0,40)`, $R^2 = 0.812$. (**d**) Low slope, high error: `x <- runif(80,0,100), y4 <- 30+1.2*x+rnorm(80,0,40)`, $R^2 = 0.429$. (**e**) Comparison of all prototype cases (a)–(d) above.

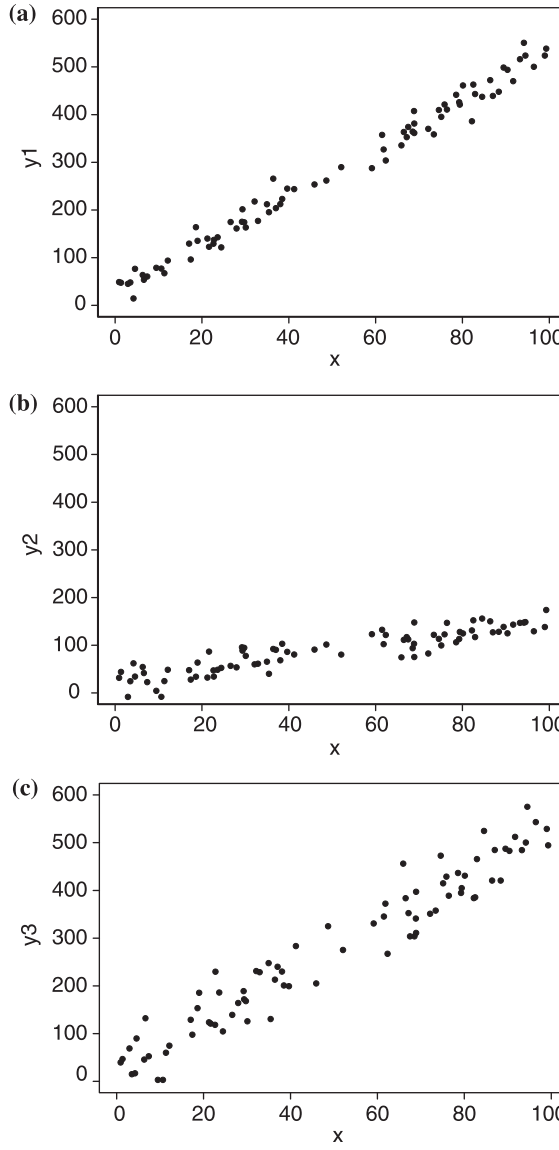
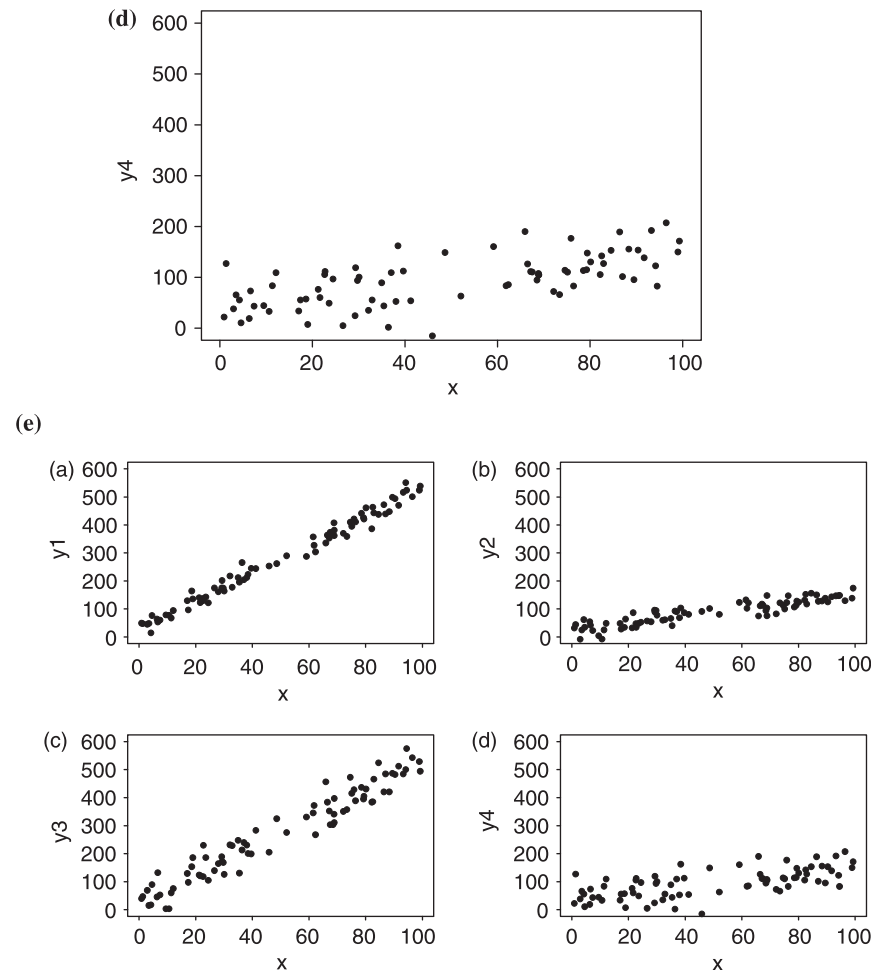


Figure A.1. *Continued.*



reduced error in fitting sample datasets. However, a model that overfits a sample dataset may not necessarily fit the population data very well. A model with $n-1$ covariates can in some cases exactly fit a dataset of sample size n. If the model more closely fits a sample dataset with additional independent covariates, the error, or bias, will be less and the R^2 value may be higher. However, the precision of the estimates of the parameters in the model may be reduced because of more parameter coefficients to estimate with the same size sample dataset. The precision of the estimates may also be reduced if the covariates are correlated with each other, thereby reducing the "stability" and increasing the error of the

estimates. So, although a higher R^2 value may in general be a worthwhile indicator of a better-fitting model, there is a danger of overfitting models using R^2 as a sole criterion for model fitting.

In conclusion, the analyst should examine the R^2 statistic but use it judiciously along with other statistics to evaluate the fit of models. Biological models with a significant amount of stochasticity may have low R^2 values, yet be well- or best-fitting, and, conversely, models with high R^2 statistics may overfit sample datasets and be overparameterized. Models with best-fitting relationships to sample datasets are not exclusively characterized by high R^2 values.

A.3.3 The Residual Standard Error $s_{y|x}$

Another important statistic useful for evaluating the fit of models with the linear regression and multiple linear regression models is the **residual standard error** $s_{y|x} = \hat{\sigma}$ is the approximately unbiased estimator of the standard deviation parameter σ for the error given by

$$s_{y|x} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-p-1)}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \sum_{j=1}^{p}\hat{\beta}_j \cdot x_j))^2}{(n-p-1)}}$$

$$= \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{(n-p-1)}},$$

where n is the sample size and p is the number of covariates in the model. Its square $s_{y|x}^2$ is the unbiased estimator of the variance σ^2. In the case of linear regression, $p = 1$ and the parameters are β_0, β_1, and σ. Since $s_{y|x}$ estimates the standard deviation of the error for the model, better-fitting models will tend to have lower residual standard errors. However, analogous to increases in R^2, this is true only up to a point. As with increases in R^2, $s_{y|x}$ tends to decrease with increasing numbers of covariates that reduce the amount of error in the fit of the model with the sample dataset. However, again, there is a danger of overfitting that must be considered in evaluating the fit of the model. So look for models that decrease $s_{y|x}$ but also examine other statistics such as the adjusted R^2, the Mallows C_p, and AIC, which penalize models with too many "independent" covariates for overfitting sample datasets.

A.3.4 The *F* Test

As with experimental data and analysis of variance (ANOVA), an F **test** can be conducted for linear regression and multiple linear regression analysis, calculating

the components of variance for the regression model, the regression (regr), error (resid), and total components, in the following table

Components of Variance	df	SS	MS	F	p Value
regr	$\mathrm{df}_{\mathrm{regr}}$	$\mathrm{SS}_{\mathrm{regr}}$	$\mathrm{MS}_{\mathrm{regr}}$	F_{s}	—
resid	$\mathrm{df}_{\mathrm{resid}}$	$\mathrm{SS}_{\mathrm{resid}}$	$\mathrm{MS}_{\mathrm{resid}}$		
Total	$\mathrm{df}_{\mathrm{total}}$	$\mathrm{SS}_{\mathrm{total}}$			

The degrees of freedom (df) formulas for the linear regression model are given by

$$\begin{aligned}\mathrm{df}_{\mathrm{regr}} &= 1,\\ \mathrm{df}_{\mathrm{resid}} &= n-2,\\ \mathrm{df}_{\mathrm{total}} &= n-1,\end{aligned}$$

where n is the sample size. Note that

$$\mathrm{df}_{\mathrm{total}} = \mathrm{df}_{\mathrm{regr}} + \mathrm{df}_{\mathrm{resid}}.$$

For multiple regression with p independent variables, the degrees of freedom formulas are given by

$$\begin{aligned}\mathrm{df}_{\mathrm{regr}} &= p,\\ \mathrm{df}_{\mathrm{resid}} &= n-p-1,\\ \mathrm{df}_{\mathrm{total}} &= n-1.\end{aligned}$$

The sums of squares (SS) formulas are given by

$$\mathrm{SS}_{\mathrm{regr}} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2,$$

$$\mathrm{SS}_{\mathrm{resid}} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

$$\mathrm{SS}_{\mathrm{total}} = \sum_{i=1}^{n}(y_i - \bar{y})^2.$$

As usual

$$\mathrm{SS}_{\mathrm{total}} = \mathrm{SS}_{\mathrm{regr}} + \mathrm{SS}_{\mathrm{resid}}.$$

The mean-squares (MS) formulas are

$$\begin{aligned}\mathrm{MS}_{\mathrm{regr}} &= \mathrm{SS}_{\mathrm{regr}}/\mathrm{df}_{\mathrm{regr}},\\ \mathrm{MS}_{\mathrm{resid}} &= \mathrm{SS}_{\mathrm{resid}}/\mathrm{df}_{\mathrm{resid}},\end{aligned}$$

and the F_s-test statistic is

$$F_s = \mathrm{MS}_{\mathrm{regr}}/\mathrm{MS}_{\mathrm{resid}}.$$

The null and alternative hypotheses are given by

$$\begin{aligned} H_0&: y_i = \beta_0 + e_i && \text{(null model); i.e., } \beta_1 = 0, \\ H_A&: y_i = \beta_0 + \beta_1 \cdot x_i + e_i && \text{(linear regression model)} \end{aligned}$$

(i.e., $\beta_1 \neq 0$). The F test may be significant for covariates even without good model fit since it assesses whether a model with covariates fits the data better than does the null model without covariates, so other statistics should also be examined. If the F-test results are insignificant; however, that is a good indication that the model with the covariate is not helpful.

A.3.5 Adjusted R^2

If R^2 is not always a reliable statistic for evaluating the competitive fit of models because of possible overfitting, what statistics will serve in its place? The **adjusted R^2 statistic** compensates for the problem of overfitting by extracting a "penalty" for the use of too many covariates. The adjusted R^2 statistic is given by the formula

$$\text{Adjusted } R^2 = R^2_{\mathrm{adj}} = 1 - \frac{n-i}{n-k} \cdot (1 - R^2),$$

where $n =$ sample size, $i = 1$ if there is intercept and 0 otherwise, and $k =$ the number of parameters. Models with the highest adjusted R^2 are the best-fitting models. It is clear from the formula that increasing numbers of covariates negatively affect the magnitude of R^2_{adj}. You can avoid the problem of models overfitting sample datasets by examining the adjusted R^2 statistic rather than the R^2 statistic. However, the adjusted R^2 statistic still may tend to overfit models to sample datasets, favoring models with too many variables. In Sections A.3.6–A.3.8, we will examine other statistics, Mallows' C_p, AIC, $\mathrm{AIC_c}$, and BIC that most effectively address the problem of overfitting.

A.3.6 Mallows' C_p

Another statistic designed to assess LS fit for models with normal residuals having constant variance is the **Mallows C_p statistic**

$$C_p = p + (n-p) \cdot \frac{(\hat{\sigma}^2 - \hat{\sigma}^2_{\mathrm{full}})}{\hat{\sigma}^2},$$

where p is the number of covariates in the model, the full model is the model with all the explanatory covariates under consideration, and

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n} = \frac{n-p-1}{n} \cdot s^2_{y|x} \cdot$$

Models with a lower Mallows C_p value are better-fitting, or more accurately, models with $C_p \cong p$ are the best-fitting models. The Mallows C_p provides approximately the same ranking for models as AIC and AIC_c (see Section A3.7). If the full model is not overfitting, then presumably models with C_p approximately equal to p will also not be overfitting. Models with C_p below p will have underestimated the error and be overfitting. Because of these rather questionable assumptions, we recommend the use of AIC and AIC_c (below) as the most rigorous and theoretically justified approach to model fitting of natural resource datasets.

A.3.7 Akaike's Information Criterion: AIC and AIC_c

We now describe the important information-theoretic statistic that is most effective for evaluating the competitiveness of models at fitting natural resource data. Akaike's information criterion was developed in the early 1970s by the Japanese mathematician Hirotugu Akaike (1973). It is an information-theoretic measurement of the **Kullback–Liebler distance** between a model and reality. **Akaike's information criterion (AIC)** and the **corrected Akaike information criterion (AIC_c)** for multiple linear regression are given by the formulas

$$\begin{aligned}
\text{AIC} &= n \cdot \log\left(s^2_{y|x} \cdot \frac{n-p-1}{n}\right) + 2 \cdot k \\
&= n \cdot \log\left(s^2_{y|x} \cdot \frac{n-k+1}{n}\right) + 2 \cdot k, \\
\text{AIC}_c &= n \cdot \log\left(s^2_{y|x} \cdot \frac{n-p-1}{n}\right) + 2 \cdot k + 2 \cdot \frac{k \cdot (k+1)}{(n-k-1)} \\
&= n \cdot \log\left(s^2_{y|x} \cdot \frac{n-k+1}{n}\right) + 2 \cdot k + 2 \cdot \frac{k \cdot (k+1)}{(n-k-1)},
\end{aligned}$$

where p is the number of covariates and $k = p + 2$ is the number of parameters (including the intercept and σ). For the linear regression model with the parameters β_0, β_1, and σ, $p = 1$ and $k = 3$. AIC is the linear Taylor series approximation of the Kullback–Leibler distance, whereas AIC_c is a second-order Taylor series approximation. Since AIC_c is more precise, it should always be used in preference to AIC, particularly for datasets with small sample size. The best-fitting model has the lowest AIC_c value.

The AIC_c criterion can be used to determine the most parsimonious model, the one with the most optimal combination of minimal bias and maximal precision. It penalizes a model with too many covariates from overfitting sample data. As we have emphasized, other traditionally popular regression statistics, such as R^2 and $s_{y|x}$, tend to favor models that overfit sample datasets, whereas AIC_c prevents such overfitting from occurring. For sound predictive models, the objective is to develop good models from sample datasets that provide reliable predictions for populations. The

objective is not merely to describe sample datasets. As the number of covariates increases, models tend to more closely fit sample datasets and reduce the bias. However, as the number of covariates increases and the number of samples in the dataset remains fixed, the precision of the covariate coefficient parameter estimates tends to decrease. The AIC_c criterion moderates this process, producing an optimal compromise between reduced bias and maximal precision. It determines the most parsimonious model, the one with the combined least amount of bias, closest to the sample dataset, and most amount of precision, with the lowest amount of sampling error, for the coefficient estimates relative to the reduced bias.

The AIC_c criterion measures the relative amount of noise, or **entropy**, in the sample data, separating it from the **signal** or **information**. It is a relative measure, since the reality is unknown; the absolute measure of entropy is the calculated AIC_c plus a constant. The constant remains unknown, but since each model has the same constant, AIC_cs may be compared to determine the best fitting models. As AIC_c provides comparative measures of fit between models only, goodness-of-fit tests must also be used in analysis to assess how well-fitting are the best-fitting models.

In general, for any probabilistic statistical model for a sample dataset with a likelihood function $\mathcal{L}$ (see Chapters 2–4, and 6, and 7), AIC and AIC_c are defined using the deviance$=D=-2\cdot\log(\mathcal{L})$:

$$\begin{aligned}
\mathrm{AIC} &= D + 2\cdot k \\
&= -2\cdot\log(\mathcal{L}) + 2\cdot k, \\
\mathrm{AIC_c} &= D + 2\cdot k + 2\cdot\frac{k\cdot(k+1)}{n-k-1} \\
&= -2\cdot\log(\mathcal{L}) + 2\cdot k + 2\cdot\frac{k\cdot(k+1)}{n-k-1}.
\end{aligned}$$

A.3.8 Bayesian Information Criterion (BIC)

The AIC_c criterion should be applied to models of realities that are complex and infinite- or high-dimensional as are most biological populations (Burnham and Anderson 1998). For such complex realities, finite-dimensional models will necessarily be inaccurate and, at best, approximations. For realities that are finite-dimensional, of fairly low dimension such as $k = 1$–5, with k fixed as the sample size n increases, so-called dimension-consistent criteria such as the Bayesian information criterion (BIC) should be applied.

The **Bayesian information criterion**, developed by Schwarz (1978), also uses a formula based on the deviance or log likelihood and "penalizes" models for the overuse of covariates

$$\begin{aligned}
\mathrm{BIC} &= D + k\cdot\log(n) \\
&= -2\cdot\log(\mathcal{L}) + k\cdot\log(n).
\end{aligned}$$

For multiple linear regression models, BIC is given by

$$\begin{aligned}\text{BIC} &= n \cdot \log\left(s^2_{y|x} \cdot \frac{n-p-1}{n}\right) + k \cdot \log(n) \\ &= n \cdot \log\left(s^2_{y|x} \cdot \frac{n-k+1}{n}\right) + k \cdot \log(n)\end{aligned}$$

and is derived using Bayesian assumptions of equal priors on each model and vague priors on the parameters (Burnham and Anderson 1998), with the objective of predicting rather than understanding the process of a system. It penalizes more heavily for increases in the number of parameters and hence sometimes tends to select models that are underfit with excessive bias. For natural resource modeling, most realities are likely complex and infinite-dimensional; hence, AIC_c is a more appropriate criteria for comparing statistical models.

A.4 STEPWISE MULTIPLE LINEAR REGRESSION METHODS

In this section we will briefly describe methods for model selection that are applicable to the multiple linear regression case where there are p covariates $\{x_1, x_2, \ldots, x_p\}$ along with the response y with n samples. We shall discuss two exploratory, descriptive, "data dredging" methods for model selection, ways of selecting from among collections of models consisting of linear combinations of covariates. For example, suppose that we are interested in developing a habitat selection model (Manly et al. 1995, 2004) using a collection of habitat covariates as predictor variables and an animal abundance or presence–absence response variable. We could use any of the following criteria as a basis for model selection:

1. Most significant coefficient estimate (i.e., lowest p value)
2. Lowest residual standard error $s_{y|x}$,
3. Highest coefficient of determination R^2
4. Highest adjusted R^2
5. Lowest Mallows C_p, or one closest to p
6. Lowest AIC
7. Lowest BIC

With **stepwise selection multiple linear regression**, we can choose a "best" model according to the covariate coefficient estimate p values, using the following iterative procedure:

1. Choose type I error bounds α_e for the entering covariate coefficients and α_s for the staying covariate coefficients with $\alpha_\text{e} < \alpha_\text{s}$.
2. Start with the null model with no covariates $y_i = \beta_0 + e_i$.

3. Consider all single-covariate models fitted to the sample dataset and select the covariate, say, x_{c_1}, with coefficient estimate $\hat{\beta}_{c_1}$ having the lowest p value below α_e, obtaining the single covariate model $y_i = \beta_0 + \beta_{c_1} \cdot x_{c_1,i} + e_i$.
4. Consider the addition of another covariate, $x_{c_{j+1}}$, to the currently selected covariate model with $x_{c_1}, x_{c_2}, \ldots, x_{c_j}$ fitted to the sample dataset and add the covariate, say, $x_{c_{j+1}}$, with coefficient estimate $\hat{\beta}_{c_{j+1}}$ having the lowest p value below α_e, if there is one, obtaining the $(k+1)$ covariate model $y_i = \beta_0 + \beta_{c_1} \cdot x_{c_1,i} + + \beta_{c_2} \cdot x_{c_2,i} + \cdots + \beta_{c_{j+1}} \cdot x_{c_{j+1},i} + e_i$; or otherwise stop.
5. Consider the p values of the current model covariate coefficient estimates and drop the covariates with p values above α_s, if there are any.
6. Continue iterating steps 4 and 5, adding and dropping covariates as appropriate until the process stops.

Note that it is important that $\alpha_e < \alpha_s$, or otherwise the process could continue indefinitely, with covariates added and dropped repeatedly.

Forward selection multiple linear regression adds covariates, without dropping any, until the process stops, whereas **backward elimination multiple linear regression** begins with the full model consisting of all covariates and drops covariates with the highest p values, without adding any, until the process stops. Stepwise selection multiple linear regression combines both forward selection and backward elimination multiple linear regression. As mentioned earlier, other optimization criteria besides p values for the coefficient estimates could be used for the covariate selection, such as lowest residual standard error $s_{y|x}$, highest R^2, highest adjusted R^2, lowest Mallows C_p, lowest AIC, or lowest BIC.

Stepwise selection, forward selection, and backward elimination multiple linear regression methods provide convenient methods for choosing models with multivariate sample datasets. However, these methods tend to overfit sample data because of the compounding of type I error caused by the multiple testing of hypotheses in the selection criteria. Therefore, these "data dredging" methods should be viewed as exploratory and descriptive. Results should be interpreted tentatively. These methods are most suitable for formulating model hypotheses that can be tested with additional sample datasets. Only with sufficient goodness-of-fit testing of the inferences, preferably with additional sample datasets, should the results of these methods be used for prediction.

A.5 BEST-SUBSETS SELECTION MULTIPLE LINEAR REGRESSION

Best-subsets selection multiple linear regression selects best-fitting models from the collection of all models that are linear combinations of covariates using some of the following criteria:

1. Highest R^2
2. Highest adjusted R^2

3. Lowest Mallows C_p, or C_p closest to p
4. Lowest AIC
5. Lowest BIC

As with stepwise selection multiple linear regression, best-subsets selection multiple regression is prone to type I error and tends to overfit sample data. As such, results should be viewed as tentative, subject to further study with additional sample datasets, if inferences are to be used for prediction.

A.6 GOODNESS OF FIT

The discussion so far has focused on the interpretation of statistics evaluating the significance and competitiveness of models at fitting sample data. The significance of the model coefficient estimates tests the null hypothesis that the coefficients are equal to 0. The significance of the F test addresses the hypothesis of whether the model differs from the null model $y_i = \beta_0 + e_i$. Other statistics, such as R^2, $s_{y|x}$, the adjusted R^2, Mallows C_p, and, most importantly, AIC_c or BIC, evaluate the fit of the model and serve as comparative statistics useful in evaluating the relative competitiveness of models at differing from the null model and fitting the sample dataset. None of these tests or statistics, however, serve adequately in evaluating model goodness of fit to the reality represented by the sample dataset.

So it is very important to examine goodness of fit of the best-fitting models as part of the overall analysis process. In the following sections, we will briefly address this issue.

A.6.1 Residual Analysis

An important part of goodness-of-fit analysis is always to examine the residuals of a linear regression or multiple linear regression model

$$\hat{e}_i = y_i - \hat{y}_i,$$

where $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j \cdot x_j$, to determine whether the assumptions of the model might be violated. Recall that the linear regression and multiple linear regression models are based on the following assumptions for the population residuals e_i:

1. Independence
2. Normality
3. Homoscedasticity (σ^2 constant)

The residuals $\hat{e}_i$ are estimates of the population residuals e_i and should be examined for apparent violations of these three assumptions. Residual plots that graph the residuals as a function of model fit are readily available in most statistical software. They should be examined for any apparent dependences among the residuals, lack of normality, or heteroscedasticity.

The residuals should occur "randomly," both positively and negatively above and below the model fit axis, without any apparent dependence relationships among them. They should also be normally distributed, with a standard deviation estimated by $s_{y|x}$. To check for normalcy, a normal plot of the residuals may be examined for linearity and a test for normality such as the Anderson–Darling test may be applied. The homoscedasticity assumption of constant variance of the residuals should be evaluated visually in the graph of the residuals, and tested, with Bartlett's test or other tests for homoscedasticity such as Levene's test or the F test. Outliers should be scrutinized for their size and quantity. If the confidence level is 95%, the proportion of outliers beyond the 95% confidence band should not greatly exceed the expected 5%. With small sample sizes, residual analysis can be more of an art than a science in practical application. So a biological explanation may provide the best rationale for the validity of the assumptions. Interested readers may consult Cook (1998) and Cook and Weisberg (1999) to examine this topic of residual analysis in more detail.

A.6.2 Confidence Intervals

Confidence intervals for the predicted mean $\hat{y}|x = \hat{\beta}_1 + \hat{\beta}_1 \cdot x$ may be calculated using the standard error for the mean given by

$$\mathrm{se}_{\hat{y}|x} = s_{y|x} \cdot \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1) \cdot s_x^2}},$$

where $s_{y|x}$ is the standard error, $\bar{x}$ is the estimated mean of the $\{x_i\}$, s_x^2 is the estimated variance of the $\{x_i\}$, and n is the sample size. Confidence intervals $\mathrm{CI}_{\hat{y}|x}$ may then be calculated and graphed using the formula

$$\mathrm{CI}_{\hat{y}|x} = \hat{y}|x \pm t_{\mathrm{d}f,1-\alpha/2} \cdot \mathrm{se}_{\hat{y}|x},$$

where the degrees of freedom $\mathrm{df} = n - k$ with k parameters in the model and $P = 1 - \alpha$ is the confidence level. The frequentist interpretation of this confidence interval with confidence level P is that, with repeated sampling, P of the sample confidence intervals will on average contain the population mean $\mu|x$ at each x.

A.6.3 Prediction Intervals

Prediction intervals for the predicted $y|x$ may also be calculated and graphed. The expected proportions of the (x_i, y_i) points in the developmental and test datasets should fall within these prediction intervals, for example, 95% of the points within the 95% prediction intervals. Prediction intervals can be calculated using the

formula for the standard error for the predicted $y|x$ given by

$$\text{se}_{y|x} = s_{y|x} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1) \cdot s_x^2}},$$

where $s_{y|x}$ is the standard error, $\bar{x}$ is the sample mean of the $\{x_i\}$, s_x^2 is the variance of the $\{x_i\}$, and n is the sample size. Note that there are three components of error in this standard error formula, indicated by the three terms in the square root, the first (i.e., 1) due to variation of the y for a fixed x, and the latter two [$1/n$ and $(x - \bar{x})^2/(n-1) \cdot s_x^2$] due to variation of the estimates for the linear regression. Using the degrees of freedom, $\text{df} = n - k$ with k parameters in the model, prediction intervals $\text{PI}_{y|x}$ with $P = 1 - \alpha$ level of confidence can be calculated:

$$\text{PI}_{y|x} = \hat{y}|x \pm t_{\text{df},1-\alpha/2} \cdot \text{se}_{y|x}.$$

A.6.4 Cross-Validation and Testing Techniques

For a linear or multiple linear regression model to provide a predictive tool, additional confirmation of the reliability of the model may be examined using either cross-validation with the developmental dataset or testing techniques with additional test datasets. If additional test datasets are available, randomly sampled from the population, goodness of fit can be examined by evaluating the predictive performance of the prediction intervals on the test datasets. For example, 95% prediction intervals can be examined on the test datasets to determine whether they perform as expected, with approximately 95% of the samples in the test dataset within the prediction interval. Alternatively, if additional test datasets are unavailable, the next best alternative is to perform cross-validation analysis on the developmental dataset. The most common method of cross-validation consists of successively omitting individual points, fitting the model to the remaining sample dataset, and examining the deleted point with respect to the prediction interval. The overall predictive accuracy of the deleted points falling within the prediction intervals is an indication of the predictive capabilities of the fitted model.