CHAPTER 3

Normal Linear Models: Statistical Inference

Chapter 2 introduced least squares fitting of ordinary linear models. For n independent observations $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$ for $\mu_i = E(y_i)$ and a model matrix $\boldsymbol{X}$ and parameter vector $\boldsymbol{\beta}$, this model states that

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta} \quad \text{with} \quad \boldsymbol{V} = \text{var}(\boldsymbol{y}) = \sigma^2\boldsymbol{I}.$$

We now add to this model the assumption that $\{y_i\}$ have normal distributions. The model is then the *normal linear model*. This chapter presents the foundations of statistical inference about the parameters of the normal linear model.

We begin this chapter by reviewing relevant distribution theory for normal linear models. Quadratic forms incorporating normally distributed response variables and projection matrices generate chi-squared distributions. One such result, *Cochran's theorem*, is the basis of significance tests about $\boldsymbol{\beta}$ in the normal linear model. Section 3.2 shows how the tests use the chi-squared quadratic forms to construct test statistics having F distributions. A useful general result about comparing two nested models is also derived as a likelihood-ratio test. Section 3.3 presents confidence intervals for elements of $\boldsymbol{\beta}$ and expected responses as well as prediction intervals for future observations. Following an example in Section 3.4, Section 3.5 presents methods for making multiple inferences with a fixed overall error rate, such as multiple comparison methods for constructing simultaneous confidence intervals for differences between all pairs of a set of means. Without the normality assumption, the exact inference methods of this chapter apply to the ordinary linear model in an approximate manner for large n.

Foundations of Linear and Generalized Linear Models, First Edition. Alan Agresti.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

3.1 DISTRIBUTION THEORY FOR NORMAL VARIATES

Statistical inference for normal linear models uses sampling distributions derived from quadratic forms with multivariate normal random variables. We now review the multivariate normal distribution and related sampling distributions.

3.1.1 Multivariate Normal Distribution

Let $N(\boldsymbol{\mu}, \boldsymbol{V})$ denote the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$. If $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ has this distribution and $\boldsymbol{V}$ is positive definite, then the probability density function (pdf) is

$$f(\boldsymbol{y}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{V}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right],$$

where $|\boldsymbol{V}|$ denotes the determinant of $\boldsymbol{V}$. Here are a few properties, when $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$.

- If $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}$, then $\boldsymbol{x} \sim N(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^{\mathrm{T}})$.
- Suppose that $\boldsymbol{y}$ partitions as

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix}, \text{ with } \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \text{ and } \boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{pmatrix}.$$

  The marginal distribution of $\boldsymbol{y}_a$ is $N(\boldsymbol{\mu}_a, \boldsymbol{V}_{aa})$, $a = 1, 2$. The conditional distribution

$$(\boldsymbol{y}_1 \mid \boldsymbol{y}_2) \sim N\left[\boldsymbol{\mu}_1 + \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2),\ \boldsymbol{V}_{11} - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{V}_{21}\right].$$

  In addition, $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are independent if and only if $\boldsymbol{V}_{12} = \boldsymbol{0}$.
- From the previous property, if $\boldsymbol{V} = \sigma^2\boldsymbol{I}$, then $y_i \sim N(\mu_i, \sigma^2)$ and $\{y_i\}$ are independent.

The normal linear model assumes that $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$ with $\boldsymbol{V} = \sigma^2\boldsymbol{I}$. The least squares estimator $\hat{\boldsymbol{\beta}}$ and the residuals $\boldsymbol{e}$ also have multivariate normal distributions, since they are linear functions of $\boldsymbol{y}$, but their elements are typically correlated. This estimator $\hat{\boldsymbol{\beta}}$ is also the maximum likelihood (ML) estimator under the normality assumption (as we showed in Section 2.1).

3.1.2 Chi-Squared, *F*, and *t* Distributions

Let χ^2_p denote a chi-squared distribution with p degrees of freedom (df). A chi-squared random variable is nonnegative with mean $= df$ and variance $= 2(df)$. Its distribution[1] is skewed to the right but becomes more bell-shaped as df increases.

[1]The pdf is the special case of the gamma distribution pdf (4.29) with shape parameter $k = df/2$.

Recall that when $y_1, \ldots, y_p$ are independent standard normal random variables, $\sum_{i=1}^{p} y_i^2 \sim \chi_p^2$. In particular, if $y \sim N(0,1)$, then $y^2 \sim \chi_1^2$. More generally

- If a p-dimensional random variable $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$ with $\boldsymbol{V}$ nonsingular of rank p, then

$$x = (\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \sim \chi_p^2.$$

Exercise 3.1 outlines a proof.

- If $z \sim N(0,1)$ and $x \sim \chi_p^2$, with x and z independent, then

$$\frac{z}{\sqrt{x/p}} \sim t_p,$$

the t distribution with $df = p$.

The t distribution is symmetric around 0 with variance $= df/(df-2)$ when $df > 2$. The term x/p in the denominator is a mean of p independent squared $N(0,1)$ random variables, so as $p \to \infty$ it converges in probability to their expected value of 1. Therefore, the t distribution converges to a $N(0,1)$ distribution as df increases.

Here is a classic way the t distribution occurs for independent responses $y_1, \ldots, y_n$ from a $N(\mu, \sigma^2)$ distribution with sample mean $\bar{y}$ and sample variance s^2: For testing H_0: $\mu = \mu_0$, the test statistic $z = \sqrt{n}(\bar{y} - \mu_0)/\sigma$ has the $N(0,1)$ null distribution. Also, s^2/σ^2 is a χ_{n-1}^2 variate $x = (n-1)s^2/\sigma^2$ divided by its df. Since $\bar{y}$ and s^2 are independent for independent observations from a normal distribution, under H_0

$$t = \frac{z}{\sqrt{x/(n-1)}} = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}.$$

Larger values of $|t|$ provide stronger evidence against H_0.

- If $x \sim \chi_p^2$ and $y \sim \chi_q^2$, with x and y independent, then

$$\frac{x/p}{y/q} \sim F_{p,q},$$

the F distribution with $df_1 = p$ and $df_2 = q$.

An F random variable takes nonnegative values. When $df_2 > 2$, it has mean $= df_2/(df_2 - 2)$, approximately 1 for large df_2. We shall use this distribution for testing hypotheses in ANOVA and regression by taking a ratio of independent mean squares. For a t random variable, t^2 has the F distribution with $df_1 = 1$ and df_2 equal to the df for that t.

3.1.3 Noncentral Distributions

In significance testing, to analyze the behavior of test statistics when null hypotheses are false, we use *noncentral* sampling distributions that occur under parameter values from the alternative hypothesis. Such distributions determine the power of a test (i.e., the probability of rejecting H_0), which can be analyzed as a function of the actual parameter value. When observations have a multivariate normal distribution, sampling distributions in such non-null cases contain the ones just summarized as special cases.

Let $\chi^2_{p,\lambda}$ denote a noncentral chi-squared distribution with $df = p$ and with noncentrality parameter λ. This is the distribution of $x = \sum_{i=1}^{p} y_i^2$ in which $\{y_i\}$ are independent with $y_i \sim N(\mu_i, 1)$ and $\lambda = \sum_{i=1}^{p} \mu_i^2$. For this distribution[2], $E(x) = p + \lambda$ and $\text{var}(x) = 2(p + 2\lambda)$. The ordinary (central) chi-squared distribution is the special case with $\lambda = 0$.

- If a p-dimensional random variable $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$ with $\boldsymbol{V}$ nonsingular of rank p, then

$$x = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{y} \sim \chi^2_{p,\lambda} \quad \text{with} \quad \lambda = \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{\mu}.$$

The construction of the noncentral chi-squared from a sum of squared independent $N(\mu_i, 1)$ random variables results when $\boldsymbol{V} = \boldsymbol{I}$.

- If $z \sim N(\mu, 1)$ and $x \sim \chi^2_p$, with x and z independent, then

$$t = \frac{z}{\sqrt{x/p}} \sim t_{p,\mu},$$

the noncentral t distribution with $df = p$ and noncentrality μ.

The noncentral t distribution is unimodal, but skewed in the direction of the sign of $\mu = E(z)$. When $p > 1$ and $\mu \neq 0$, its mean $E(t) \approx [1 - 3/(4p - 1)]^{-1}\mu$, which is near μ but slightly larger in absolute value. For large p, the distribution of t is approximately the $N(\mu, 1)$ distribution.

- If $x \sim \chi^2_{p,\lambda}$ and $y \sim \chi^2_q$, with x and y independent, then

$$\frac{x/p}{y/q} \sim F_{p,q,\lambda},$$

the noncentral F distribution with $df_1 = p$, $df_2 = q$, and noncentrality λ.

[2]Here is an alternative way to define noncentrality: Let $z \sim \text{Poisson}(\phi)$ and $(x \mid z) \sim \chi^2_{p+2z}$. Then unconditionally $x \sim \chi^2_{p,\phi}$. This noncentrality ϕ relates to the noncentrality λ we defined by $\phi = \lambda/2$.

For large df_2, the noncentral F has mean approximately $1 + \lambda/df_1$, which increases in λ from the approximate mean of 1 for the central case.

As reality deviates farther from a particular null hypothesis, the noncentrality λ increases. The noncentral chi-squared and noncentral F distributions are stochastically increasing in λ. That is, evaluated at any positive value, the cumulative distribution function (cdf) decreases as λ increases, so values of the statistic tend to be larger.

3.1.4 Normal Quadratic Forms with Projection Matrices Are Chi-Squared

Two results about quadratic forms involving normal random variables are especially useful for statistical inference with normal linear models. The first generalizes the above quadratic form result for the noncentral chi-squared, which follows with $\boldsymbol{A} = \boldsymbol{V}^{-1}$.

- Suppose $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$ and $\boldsymbol{A}$ is a symmetric matrix. Then,

$$\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{y} \sim \chi^2_{r,\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu}} \quad \Leftrightarrow \quad \boldsymbol{A}\boldsymbol{V} \text{ is idempotent of rank } r.$$

For the normal linear model, the n independent observations $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ with $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, and so $\boldsymbol{y}/\sigma \sim N(\boldsymbol{\mu}/\sigma, \boldsymbol{I})$. By this result, if $\boldsymbol{P}$ is a projection matrix (which is symmetric and idempotent) with rank r, then $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{y}/\sigma^2 \sim \chi^2_{r,\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{\mu}/\sigma^2}$. Applying the result with the standardized normal variables $(\boldsymbol{y} - \boldsymbol{\mu})/\sigma \sim N(\boldsymbol{0}, \boldsymbol{I})$, we have

Normal quadratic form with projection matrix and chi-squared: Suppose $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ and $\boldsymbol{P}$ is symmetric. Then,

$$\frac{1}{\sigma^2}(\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{P}(\boldsymbol{y} - \boldsymbol{\mu}) \sim \chi^2_r \quad \Leftrightarrow \quad \boldsymbol{P} \text{ is a projection matrix of rank } r.$$

Cochran (1934) showed[3] this result, which also provides an interpretation for degrees of freedom.

- Since the *df* for the chi-squared distribution of a quadratic form with a normal linear model equals the rank of $\boldsymbol{P}$, *degrees of freedom* represent the dimension of the vector subspace to which $\boldsymbol{P}$ projects.

The following key result also follows from Cochran (1934), building on the first result.

[3]From Cochran's result I, since a symmetric matrix whose eigenvalues are 0 and 1 is idempotent.

Cochran's theorem: Suppose n observations $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ and $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_k$ are projection matrices having $\sum_i \boldsymbol{P}_i = \boldsymbol{I}$. Then, $\{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{y}\}$ are independent and $\left(\frac{1}{\sigma^2}\right)\boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{y} \sim \chi^2_{r_i,\lambda_i}$ where $r_i = \mathrm{rank}(\boldsymbol{P}_i)$ and $\lambda_i = \frac{1}{\sigma^2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{\mu}$, $i = 1, \ldots, k$, with $\sum_i r_i = n$.

If we replace $\boldsymbol{y}$ by $(\boldsymbol{y} - \boldsymbol{\mu})$ in the quadratic forms, we obtain central chi-squared distributions ($\lambda_i = 0$). This result is the basis of significance tests for parameters in normal linear models. The proof of the independence result shows that all pairs of projection matrices in this decomposition satisfy $\boldsymbol{P}_i\boldsymbol{P}_j = 0$.

3.1.5 Proof of Cochran's Theorem

We next show a proof[4] of Cochran's theorem. You may wish to skip these technical details for now and go to the next section, which uses this result to construct significance tests for the normal linear model.

We first show that if $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ and $\boldsymbol{P}$ is a projection matrix having rank r, then $\left(\frac{1}{\sigma^2}\right)\boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{y} \sim \chi^2_{r,\lambda}$ with $\lambda = \frac{1}{\sigma^2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{\mu}$. Since $\boldsymbol{P}$ is symmetric and idempotent with rank r, its eigenvalues are 1 (r times) and 0 ($n - r$ times). By the spectral decomposition of a symmetric matrix, we can express $\boldsymbol{P} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathrm{T}}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix of $(1, 1, \ldots, 1, 0, \ldots, 0)$, the eigenvalues of $\boldsymbol{P}$, and $\boldsymbol{Q}$ is an orthogonal matrix with columns that are the eigenvectors of $\boldsymbol{P}$. Let $\boldsymbol{z} = \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{y}/\sigma$. Then, $\boldsymbol{z} \sim N(\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{\mu}/\sigma, \boldsymbol{I})$, and $\left(\frac{1}{\sigma^2}\right)\boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{y} = \boldsymbol{z}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{z} = \sum_{i=1}^{r} z_i^2$. Since each z_i is normal with standard deviation 1, $\sum_{i=1}^{r} z_i^2$ has a noncentral chi-squared distribution with $df = r$ and noncentrality parameter

$$\sum_{i=1}^{r}[E(z_i)]^2 = [E(\boldsymbol{\Lambda}\boldsymbol{z})]^{\mathrm{T}}[E(\boldsymbol{\Lambda}\boldsymbol{z})] = \left(\frac{1}{\sigma^2}\right)[\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{\mu}]^{\mathrm{T}}[\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{\mu}]$$
$$= \left(\frac{1}{\sigma^2}\right)\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{\mu} = \left(\frac{1}{\sigma^2}\right)\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{\mu}.$$

Now we consider k quadratic forms with k projection matrices that are a decomposition of $\boldsymbol{I}$, the $n \times n$ identity matrix. The rank of a projection matrix is its trace, so $\sum_i r_i = \sum_i \mathrm{trace}(\boldsymbol{P}_i) = \mathrm{trace}(\sum_i \boldsymbol{P}_i) = \mathrm{trace}(\boldsymbol{I}) = n$. We apply the spectral decomposition to each projection matrix, with $\boldsymbol{P}_i = \boldsymbol{Q}_i\boldsymbol{\Lambda}_i\boldsymbol{Q}_i^{\mathrm{T}}$, where $\boldsymbol{\Lambda}_i$ is a diagonal matrix of $(1, 1, \ldots, 1, 0, \ldots, 0)$ with r_i entries that are 1. By the form of $\boldsymbol{\Lambda}_i$, this is identical to $\boldsymbol{P}_i = \tilde{\boldsymbol{Q}}_i\boldsymbol{I}_{r_i}\tilde{\boldsymbol{Q}}_i^{\mathrm{T}} = \tilde{\boldsymbol{Q}}_i\tilde{\boldsymbol{Q}}_i^{\mathrm{T}}$, where $\tilde{\boldsymbol{Q}}_i$ is a $n \times r_i$ matrix of the first r_i columns of $\boldsymbol{Q}_i$. Note that $\tilde{\boldsymbol{Q}}_i^{\mathrm{T}}\tilde{\boldsymbol{Q}}_i = \boldsymbol{I}_{r_i}$. We stack the $\{\tilde{\boldsymbol{Q}}_i\}$ together as

$$\boldsymbol{Q} = [\tilde{\boldsymbol{Q}}_1 \ : \ \tilde{\boldsymbol{Q}}_2 \ : \ \cdots \ : \ \tilde{\boldsymbol{Q}}_k],$$

[4]This proof is based on one in Monahan (2008, pp. 113–114).

for which

$$\boldsymbol{Q}\boldsymbol{Q}^{\mathrm{T}} = \tilde{\boldsymbol{Q}}_1\tilde{\boldsymbol{Q}}_1^{\mathrm{T}} + \cdots + \tilde{\boldsymbol{Q}}_k\tilde{\boldsymbol{Q}}_k^{\mathrm{T}} = \boldsymbol{P}_1 + \cdots + \boldsymbol{P}_k = \boldsymbol{I}_n.$$

Thus, $\boldsymbol{Q}$ is an orthogonal $n \times n$ matrix and also $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q} = \boldsymbol{I}_n$ and $\tilde{\boldsymbol{Q}}_i^{\mathrm{T}}\tilde{\boldsymbol{Q}}_j = \boldsymbol{0}$ for $i \neq j$. So $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{y} \sim N(\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$, and its components $\{\tilde{\boldsymbol{Q}}_i^{\mathrm{T}}\boldsymbol{y}\}$ are independent, as are $\{\|\tilde{\boldsymbol{Q}}_i^{\mathrm{T}}\boldsymbol{y}\|^2 = \boldsymbol{y}^{\mathrm{T}}\tilde{\boldsymbol{Q}}_i\tilde{\boldsymbol{Q}}_i^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{y}\}$. Note[5] also that for $i \neq j$, $\boldsymbol{P}_i\boldsymbol{P}_j = \tilde{\boldsymbol{Q}}_i\tilde{\boldsymbol{Q}}_i^{\mathrm{T}}\tilde{\boldsymbol{Q}}_j\tilde{\boldsymbol{Q}}_j^{\mathrm{T}} = \boldsymbol{0}$.

3.2 SIGNIFICANCE TESTS FOR NORMAL LINEAR MODELS

We now use Cochran's theorem to derive fundamental significance tests for the normal linear model. We first revisit the one-way layout and then present inference for the more general context of comparing two nested normal linear models.

3.2.1 Example: ANOVA for the One-Way Layout

For the one-way layout (introduced in Sections 1.3.3 and 2.3.2), let y_{ij} denote observation j in group i, for $i = 1, \ldots, c$ and $j = 1, \ldots, n_i$, with $n = \sum_i n_i$. The observations are assumed to be independent. The linear predictor for $\mu_i = E(y_{ij})$ is

$$E(y_{ij}) = \beta_0 + \beta_i,$$

with a constraint such as $\beta_1 = 0$. We construct a significance test of H_0: $\mu_1 = \cdots = \mu_c$, assuming that $\{y_{ij} \sim N(\mu_i, \sigma^2)\}$. Under H_0, which is equivalently H_0: $\beta_1 = \cdots = \beta_c$, the model simplifies to the null model, $E(y_{ij}) = \beta_0$ for all i and j.

The projection matrix $\boldsymbol{P}_{\boldsymbol{X}}$ for this model is a block-diagonal matrix with components $\frac{1}{n_i}\mathbf{1}_{n_i}\mathbf{1}_{n_i}^{\mathrm{T}}$, shown in Equation 2.6 of Section 2.3.2. Let $\boldsymbol{P}_0 = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$ denote the projection matrix for the null model. We use the decomposition

$$\boldsymbol{I} = \boldsymbol{P}_0 + (\boldsymbol{P}_{\boldsymbol{X}} - \boldsymbol{P}_0) + (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}).$$

Each of the three components is a projection matrix, so we can apply Cochran's theorem with $\boldsymbol{P}_1 = \boldsymbol{P}_0$, $\boldsymbol{P}_2 = \boldsymbol{P}_{\boldsymbol{X}} - \boldsymbol{P}_0$, and $\boldsymbol{P}_3 = \boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}$. The ranks of the components, which equal their traces, are 1, $c - 1$, and $n - c$.

From Section 2.3.3, the corrected total sum of squares (TSS) decomposes into two parts,

$$\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{P}_{\boldsymbol{X}} - \boldsymbol{P}_0)\boldsymbol{y} = \sum_{i=1}^{c} n_i(\bar{y}_i - \bar{y})^2, \quad \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y} = \sum_{i=1}^{c}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2,$$

[5]The result that $\boldsymbol{P}_i\boldsymbol{P}_j = \boldsymbol{0}$ is also a special case of the stronger result about the decomposition of projection matrices stated at the end of Section 2.1.1.

the "between-groups" and "within-groups" sums of squares. By Cochran's theorem,

$$\frac{1}{\sigma^2}\sum_{i=1}^{c} n_i(\bar{y}_i - \bar{y})^2 \sim \chi^2_{c-1,\lambda}, \quad \text{with} \quad \lambda = \frac{1}{\sigma^2}\boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{P}_{\boldsymbol{X}} - \boldsymbol{P}_0)\boldsymbol{\mu},$$
$$\frac{1}{\sigma^2}\left[\sum_{i=1}^{c}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2\right] \sim \chi^2_{n-c},$$

and the quadratic forms are independent. The second one has noncentrality 0 because $\boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{\mu} = \boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{\mu} - \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{\mu}) = \boldsymbol{\mu}^{\mathrm{T}}\mathbf{0} = 0$. As a consequence, the test statistic

$$F = \frac{\sum_{i=1}^{c} n_i(\bar{y}_i - \bar{y})^2/(c-1)}{\sum_{i=1}^{c}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2/(n-c)} \sim F_{c-1,n-c,\lambda}.$$

Using the expressions for $\boldsymbol{P}_0$ and $\boldsymbol{P}_{\boldsymbol{X}}$, you can verify that $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{\mu} = \sum_{i=1}^{c} n_i\mu_i^2$ and $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{P}_0\boldsymbol{\mu} = n\bar{\mu}^2$, where $\bar{\mu} = \sum_i n_i\mu_i/n$. Thus, the noncentrality simplifies to $\lambda = \frac{1}{\sigma^2}\sum_{i=1}^{c} n_i(\mu_i - \bar{\mu})^2$. Under H_0, $\lambda = 0$, and the F test statistic has an F distribution with $df_1 = c - 1$ and $df_2 = n - c$. Larger F values are more contradictory to H_0, so the P-value is the right-tail probability from that distribution above the observed test statistic value, F_{obs}. When H_0 is false, λ and the power of the test increase as $\{n_i\}$ increase and as the variability in $\{\mu_i\}$ increases.

This significance test for the one-way layout is known as *(one-way) analysis of variance*, due to R. A. Fisher (1925). To complete the ANOVA table shown in Table 2.1, we include mean squares, which are ratios of the two SS values to their *df* values, and the F statistic as the ratio of those mean squares. The table has the form shown in Table 3.1, and would also include the P-value, $P_{H_0}(F > F_{\text{obs}})$. The first line refers to the null model, which specifies a common mean for all groups. Often, the ANOVA table does not show this line, essentially assuming the intercept is in the model. The table then shows the total sum of squares after subtracting $n\bar{y}^2$, giving the corrected total sum of squares, $\text{TSS} = \sum_i\sum_j(y_{ij} - \bar{y})^2$ based on $df = n - 1$.

Table 3.1 Complete ANOVA Table for the Normal Linear Model for the One-Way Layout

Source	*df*	Sum of Squares	Mean Square	F_{obs}
Mean	1	$n\bar{y}^2$		
Group	$c-1$	$\sum_i n_i(\bar{y}_i - \bar{y})^2$	$\frac{\sum_i n_i(\bar{y}_i-\bar{y})^2}{c-1}$	$\frac{\sum_i n_i(\bar{y}_i-\bar{y})^2/(c-1)}{\sum_i\sum_j(y_{ij}-\bar{y}_i)^2/(n-c)}$
Error	$n-c$	$\sum_i\sum_j(y_{ij} - \bar{y}_i)^2$	$\frac{\sum_i\sum_j(y_{ij}-\bar{y}_i)^2}{n-c}$	
Total	n	$\sum_{i=1}^{c}\sum_{j=1}^{n_i} y_{ij}^2$		

3.2.2 Comparing Two Nested Normal Linear Models

The model-building process often deals with comparing a model to a more complex one that has additional parameters or to a simpler one that has fewer parameters. An example of the first type is analyzing whether to add interaction terms to a model containing only main effects. An example of the second type is testing whether sufficiently strong evidence exists to keep a term in the model. Denote the simpler model by M_0 and the more complex model by M_1. Denote the numbers of parameters by p_0 for M_0 and p_1 for M_1, when both model matrices have full rank. We now construct a test of the null hypothesis that M_0 holds against the alternative hypothesis that M_1 holds.

Denote the projection matrices for the two models by $\boldsymbol{P}_0$ and $\boldsymbol{P}_1$. The decomposition using projection matrices

$$\boldsymbol{I} = \boldsymbol{P}_0 + (\boldsymbol{P}_1 - \boldsymbol{P}_0) + (\boldsymbol{I} - \boldsymbol{P}_1)$$

corresponds to the orthogonal decomposition of the data as

$$\boldsymbol{y} = \boldsymbol{P}_0\boldsymbol{y} + (\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{y} + (\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y}.$$

Here $\boldsymbol{P}_0\boldsymbol{y} = \hat{\boldsymbol{\mu}}_0$ and $\boldsymbol{P}_1\boldsymbol{y} = \hat{\boldsymbol{\mu}}_1$ are the fitted values for the two models. The corresponding sum-of-squares decomposition is

$$\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}_0\boldsymbol{y} + \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{y} + \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y}.$$

From Sections 2.4.1 and 2.4.2, $\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y} = \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1)^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y} = \sum_i (y_i - \hat{\mu}_{i1})^2$ is the residual sum of squares for M_1, which we denote by SSE_1. Likewise,

$$\begin{aligned}\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{y} &= \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_0)\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y} \\ &= \sum_i (y_i - \hat{\mu}_{i0})^2 - \sum_i (y_i - \hat{\mu}_{i1})^2 = \mathrm{SSE}_0 - \mathrm{SSE}_1.\end{aligned}$$

Since $(\boldsymbol{P}_1 - \boldsymbol{P}_0)$ is a projection matrix, this difference also equals

$$\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{y} = \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{y} = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^{\mathrm{T}}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0).$$

So $\mathrm{SSE}_0 - \mathrm{SSE}_1 = \sum_i (\hat{\mu}_{i1} - \hat{\mu}_{i0})^2 = \mathrm{SSR}(M_1 \mid M_0)$, the difference between the regression SS values for M_1 and M_0.

Now $\boldsymbol{I} - \boldsymbol{P}_1$ has rank $n - p_1$, since $\mathrm{trace}(\boldsymbol{I} - \boldsymbol{P}_1) = \mathrm{trace}(\boldsymbol{I}) - \mathrm{trace}(\boldsymbol{P}_1)$ and $\boldsymbol{P}_1$ has full rank p_1. Likewise, $\boldsymbol{P}_1 - \boldsymbol{P}_0$ has rank $p_1 - p_0$. Under H_0, by Cochran's theorem,

$$\frac{\mathrm{SSE}_0 - \mathrm{SSE}_1}{\sigma^2} \sim \chi^2_{p_1 - p_0} \quad \text{and} \quad \frac{\mathrm{SSE}_1}{\sigma^2} \sim \chi^2_{n - p_1},$$

and these are independent. Here, under H_0, the noncentralities of the two chi-squared variates are

$$\boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{\mu} = 0, \quad \boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{\mu} = 0$$

since for $\boldsymbol{\mu}$ satisfying M_0, $\boldsymbol{P}_1\boldsymbol{\mu} = \boldsymbol{P}_0\boldsymbol{\mu} = \boldsymbol{\mu}$. It follows that, under H_0, the test statistic

$$F = \frac{(\mathrm{SSE}_0 - \mathrm{SSE}_1)/(p_1 - p_0)}{\mathrm{SSE}_1/(n - p_1)} \tag{3.1}$$

has an F distribution with $df_1 = p_1 - p_0$ and $df_2 = n - p_1$. The denominator $\mathrm{SSE}_1/(n - p_1)$ is the error mean square, which is the s^2 estimator of σ^2 for M_1. Larger differences in SSE values, and larger values of the F test statistic, provide stronger evidence against H_0. The P-value is $P_{H_0}(F > F_{\mathrm{obs}})$.

3.2.3 Likelihood-Ratio Test Comparing Models

The test comparing two nested normal linear models can also be derived as a likelihood-ratio test[6]. For the normal linear model with model matrix $\boldsymbol{X}$, the likelihood function is

$$\ell(\boldsymbol{\beta}, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-(1/2\sigma^2)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right].$$

The log-likelihood function is

$$L(\boldsymbol{\beta}, \sigma) = -(n/2)\log(2\pi) - n\log(\sigma) - (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/2\sigma^2.$$

From Section 2.1.1, differentiating with respect to $\boldsymbol{\beta}$ yields the normal equations and the least squares estimate, $\hat{\boldsymbol{\beta}}$. Differentiating with respect to σ yields

$$\partial L(\boldsymbol{\beta}, \sigma)/\partial\sigma = -\frac{n}{\sigma} + \frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{\sigma^3}.$$

Setting this equal to 0 and solving yields the ML estimator

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n} = \frac{\mathrm{SSE}}{n}.$$

This estimator is the multiple $(n - p)/n$ of the unbiased estimator, which is $s^2 = \left[\sum_i (y_i - \hat{\mu}_i)^2\right]/(n - p)$. The maximized likelihood function simplifies to

$$\ell(\hat{\boldsymbol{\beta}}, \hat{\sigma}) = \left(\frac{1}{\hat{\sigma}\sqrt{2\pi}}\right)^n e^{-n/2}.$$

[6]The likelihood-ratio test is introduced in a more general context, for GLMs, in Section 4.3.1.

Now, for testing M_0 against M_1, let $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$ denote the two ML variance estimates. The ratio of the maximized likelihood functions is

$$\frac{\sup_{M_0} \ell(\boldsymbol{\beta}, \sigma)}{\sup_{M_1} \ell(\boldsymbol{\beta}, \sigma)} = \left(\frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2}\right)^{n/2}$$
$$= \left(\frac{\text{SSE}_1}{\text{SSE}_0}\right)^{n/2} = \left(1 + \frac{\text{SSE}_0 - \text{SSE}_1}{\text{SSE}_1}\right)^{-n/2} = \left(1 + \frac{p_1 - p_0}{n - p_1} F\right)^{-n/2}$$

for the F test statistic (3.1) derived above. A small value of the likelihood ratio, and thus strong evidence against H_0, corresponds to a large value of the F statistic.

3.2.4 Example: Test That All Effects in a Normal Linear Model Equal Zero

In an important special case of the test comparing two nested normal linear models, the simpler model M_0 is the null model, $E(y_i) = \beta_0$, and M_1 has a set of explanatory variables[7],

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1}.$$

Comparing the models corresponds to testing the global null hypothesis H_0: $\beta_1 = \cdots = \beta_{p-1} = 0$.

The projection matrix for M_0 is $\boldsymbol{P}_0 = \frac{1}{n}\mathbf{1}\mathbf{1}^{\text{T}}$. The sum-of-squares decomposition corresponding to the orthogonal decomposition

$$\boldsymbol{y} = \boldsymbol{P}_0\boldsymbol{y} + (\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{y} + (\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y}$$

yields the ANOVA table shown in Table 3.2, where $\{\hat{\mu}_i\}$ are the fitted values for the full model. The F test statistic, which is the ratio of the mean squares, has $df_1 = p - 1$ and $df_2 = n - p$.

Table 3.2 ANOVA Table for Testing That All Effects in a Normal Linear Model Equal Zero

Source	Projection Matrix	df	Sum of Squares	Mean Square
Intercept	$\boldsymbol{P}_0 = \frac{1}{n}\mathbf{1}\mathbf{1}^{\text{T}}$	1	$\boldsymbol{y}^{\text{T}}\boldsymbol{P}_0\boldsymbol{y} = n\bar{y}^2$	
Regression	$\boldsymbol{P}_1 - \boldsymbol{P}_0$	$p-1$	$\boldsymbol{y}^{\text{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{y} = \sum_i(\hat{\mu}_i - \bar{y})^2$	$\frac{\sum_i(\hat{\mu}_i - \bar{y})^2}{p-1}$
Error	$\boldsymbol{I} - \boldsymbol{P}_1$	$n-p$	$\boldsymbol{y}^{\text{T}}(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y} = \sum_i(y_i - \hat{\mu}_i)^2$	$\frac{\sum_i(y_i - \hat{\mu}_i)^2}{n-p}$
Total	$\boldsymbol{I}$	n	$\sum_{i=1}^n y_i^2$	

[7]We use $p - 1$ for the highest index, so p is, as usual, the number of model parameters.

The one-way ANOVA test for c means constructed in Section 3.2.1 results when $p = c$ and the explanatory variables are indicator variables for $c - 1$ of the c groups. Testing H_0: $\beta_1 = \cdots = \beta_{c-1} = 0$ is then equivalent to testing H_0: $\mu_1 = \cdots = \mu_c$. The fitted value $\hat{\mu}_{ij}$ is then $\bar{y}_i$.

3.2.5 Non-null Behavior of F Statistic Comparing Nested Models

The numerator of the F test statistic for comparing two models summarizes the sample information about how much better M_1 fits than M_0. A relatively large value for $\text{SSE}_0 - \text{SSE}_1 = \|\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0\|^2$ yields a large F value. If M_1 holds but M_0 does not, how large can we expect $\|\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0\|^2$ and the F test statistic to be?

When M_1 holds, $E(\boldsymbol{y}) = \boldsymbol{\mu}_1$. Since $(\boldsymbol{P}_1 - \boldsymbol{P}_0)$ is symmetric and idempotent,

$$E\|\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0\|^2 = E\|(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{y}\|^2 = E[\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{y}].$$

Using the result (2.7) shown in Section 2.4.1 for $\boldsymbol{V} = \text{var}(\boldsymbol{y})$ and a matrix $\boldsymbol{A}$ that $E(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{y}) = \text{trace}(\boldsymbol{A}\boldsymbol{V}) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu}$, we have (with $\boldsymbol{V} = \sigma^2\boldsymbol{I}$)

$$\begin{aligned} E[\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{y}] &= \text{trace}[(\boldsymbol{P}_1 - \boldsymbol{P}_0)\sigma^2\boldsymbol{I}] + \boldsymbol{\mu}_1^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{\mu}_1 \\ &= \sigma^2[\text{rank}(\boldsymbol{P}_1) - \text{rank}(\boldsymbol{P}_0)] + \boldsymbol{\mu}_1^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{\mu}_1. \end{aligned}$$

Let $\boldsymbol{\mu}_0 = \boldsymbol{P}_0\boldsymbol{\mu}_1$ denote the projection of the true mean vector onto the model space for M_0. Then, with full-rank model matrices, the numerator of the F test statistic has expected value

$$E\left[\frac{\|\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0\|^2}{p_1 - p_0}\right] = \sigma^2 + \frac{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|^2}{p_1 - p_0}.$$

The chi-squared component of the numerator of the F statistic is

$$\frac{\|\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0\|^2}{\sigma^2} \sim \chi^2_{p_1 - p_0, \lambda}$$

with noncentrality $\lambda = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|^2/\sigma^2$.

Next, for this non-null case, consider the denominator of the F statistic, which is the estimate of the error variance σ^2 for model M_1. Since

$$E\|\boldsymbol{y} - \hat{\boldsymbol{\mu}}_1\|^2 = E[\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y}] = \text{trace}[(\boldsymbol{I} - \boldsymbol{P}_1)\sigma^2\boldsymbol{I}] + \boldsymbol{\mu}_1^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{\mu}_1$$

and since $(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{\mu}_1 = \boldsymbol{0}$, this expected sum of squares equals $(n - p_1)\sigma^2$. Thus, regardless of whether H_0 is true, the F denominator has

$$E\left[\frac{\|\boldsymbol{y} - \hat{\boldsymbol{\mu}}_1\|^2}{n - p_1}\right] = \sigma^2.$$

Under H_0 for testing M_0 against M_1, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0$ and the expected value of the numerator mean square is also σ^2. Then the F test statistic is a ratio of two unbiased estimators of σ^2. The ratio of expectations equals 1, and when $n - p_1$ (and hence df_2) is large, this is also the approximate expected value of the F test statistic itself. That is, under H_0 we expect to observe F values near 1, within limits of sampling variability. Under the alternative, the ratio of the expected value of the numerator to the expected value of the denominator is $1 + \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|^2/(p_1 - p_0)\sigma^2$. The noncentrality of the F test is the noncentrality of the numerator chi-squared, $\lambda = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|^2/\sigma^2$. The power of the F test increases as n increases, since then $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ contain more elements that contribute to the numerator sum of squares in λ.

3.2.6 Expected Mean Squares and Power for One-Way ANOVA

To illustrate expected non-null behavior, consider the one-way ANOVA F test for c groups, derived in Section 3.2.1. For it, the expected value of the numerator mean square is

$$E\left[\frac{\|\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0\|^2}{p_1 - p_0}\right] = E\left[\frac{\sum_{i=1}^{c} n_i(\bar{y}_i - \bar{y})^2}{c-1}\right] = \sigma^2 + \frac{\sum_{i=1}^{c} n_i(\mu_i - \bar{\mu})^2}{c-1}.$$

Suppose the ANOVA compares $c = 3$ groups with $n_i = 10$ observations per group. The F test statistic for H_0: $\mu_1 = \mu_2 = \mu_3$ has $df_1 = 2$ and $df_2 = n - 3 = 27$. Let $F_{q,a,b}$ denote the q quantile of the central F distribution with $df_1 = a$ and $df_2 = b$. Consider the relatively large effects $\mu_1 - \mu_2 = \mu_2 - \mu_3 = \sigma$. The noncentrality (derived in Section 3.2.1) of $\lambda = \frac{1}{\sigma^2}\sum_i n_i(\mu_i - \bar{\mu})^2$ then equals 20. The power of the F test with size $\alpha = 0.05$ is the probability that a noncentral F random variable with $df_1 = 2, df_2 = 27$, and $\lambda = 20$ exceeds $F_{0.95,2,27}$. Using R, we find that the power is quite high, 0.973:

```
------------------------------------------------------------------------
> qf(0.95, 2, 27) # 0.95 quantile of F dist. with df1 = 2, df2 = 27
[1] 3.354131
> 1 - pf(3.354131, 2, 27, 20) # right-tail prob. for noncentral F
[1] 0.9732551
------------------------------------------------------------------------
```

In planning a study, it is sensible to find the power for various n for a variety of plausible effect sizes.

3.2.7 Testing a General Linear Hypothesis

In practice, nearly all hypotheses tested about effects in linear models can be expressed as H_0: $\boldsymbol{\Lambda\beta} = \mathbf{0}$ for a $\ell \times p$ matrix of constants $\boldsymbol{\Lambda}$ and a vector of estimable quantities $\boldsymbol{\Lambda\beta}$. A special case is the example just considered of H_0: $\beta_1 = \cdots = \beta_{p-1} = 0$ for comparing a full model to the null model. Another example is a test for a contrast or set of contrasts, such as H_0: $\beta_j - \beta_k = 0$ for comparing means j and k in a

one-way layout (see Section 3.4.5). The form H_0: $\boldsymbol{\Lambda\beta} = \mathbf{0}$ is called the *general linear hypothesis*.

Suppose $\boldsymbol{X}$ and $\boldsymbol{\Lambda}$ are full rank, so the hypotheses contain no redundancies. That is, H_0 imposes ℓ independent constraints on an identifiable $\boldsymbol{\beta}$. The estimator $\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\Lambda\beta}$ is the BLUE, and it is maximum likelihood under the assumption of normality for y. As a vector of ℓ linear transformations of $\hat{\boldsymbol{\beta}}$, $\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}$ has a $N[\boldsymbol{\Lambda\beta}, \boldsymbol{\Lambda}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{\Lambda}^{\mathrm{T}}\sigma^2]$ distribution. The quadratic form

$$(\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}} - \mathbf{0})^{\mathrm{T}} \left[\boldsymbol{\Lambda}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{\Lambda}^{\mathrm{T}}\sigma^2\right]^{-1} (\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}} - \mathbf{0})$$

compares the estimate $\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\Lambda\beta}$ to its H_0 value of $\mathbf{0}$, relative to the inverse covariance matrix of $\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}$. Under H_0, it has a chi-squared distribution with $df = \ell$. By the orthogonality of the model space and the error space, we can form an F test statistic (with $df_1 = \ell$ and $df_2 = n - p$) from the ratio of chi-squared variates divided by their df values,

$$F = \frac{(\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}})^{\mathrm{T}} \left[\boldsymbol{\Lambda}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{\Lambda}^{\mathrm{T}}\right]^{-1} \boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}/\ell}{\mathrm{SSE}/(n-p)},$$

where σ^2 has canceled from the numerator and denominator.

The restriction $\boldsymbol{\Lambda\beta} = \mathbf{0}$ implies a new model that is a special case M_0 of the original model. In fact, the F statistic just derived is identical to the F statistic (3.1) for comparing the full model to the special case M_0. So, how can we express the original model and the constraints $\boldsymbol{\Lambda\beta} = \mathbf{0}$ as an equivalent model M_0? It is the model having model matrix $\boldsymbol{X}_0$ found as follows. Let $\boldsymbol{U}$ be a matrix such that $C(\boldsymbol{U})$ is the orthogonal complement of $C(\boldsymbol{\Lambda}^T)$. That is, $\boldsymbol{\beta}$ is such that $\boldsymbol{\Lambda\beta} = \mathbf{0}$ if and only if $\boldsymbol{\beta} \in C(\boldsymbol{U})$. Then $\boldsymbol{\beta} = \boldsymbol{U\gamma}$ for some vector $\boldsymbol{\gamma}$. But under this restriction the original model $E(\boldsymbol{y}) = \boldsymbol{X\beta}$ simplifies to $E(\boldsymbol{y}) = \boldsymbol{XU\gamma} = \boldsymbol{X}_0\boldsymbol{\gamma}$ for $\boldsymbol{X}_0 = \boldsymbol{XU}$. Also, M_0 is a simpler model than the original model, with $C(\boldsymbol{X}_0)$ contained in $C(\boldsymbol{X})$, since any vector that is a linear combination of columns of $\boldsymbol{X}_0$ (e.g., $\boldsymbol{X}_0\boldsymbol{\gamma}$) is also a linear combination of columns of $\boldsymbol{X}$ (e.g., $\boldsymbol{X\beta}$ with $\boldsymbol{\beta} = \boldsymbol{U\gamma}$).

In the F statistic for comparing the two models, it can be shown[8] that

$$\mathrm{SSE}_0 - \mathrm{SSE}_1 = (\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}})^{\mathrm{T}} \left[\boldsymbol{\Lambda}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{\Lambda}^{\mathrm{T}}\right]^{-1} \boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}.$$

When we developed the F test for comparing nested models in Section 3.2.2, we observed that $\mathrm{SSE}_0 - \mathrm{SSE}_1$ was merely $\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{y}$ based on the projection matrices for the two models. For the general linear hypothesis, what is the difference $(\boldsymbol{P}_1 - \boldsymbol{P}_0)$ projection matrix? Using the least squares solution for $\hat{\boldsymbol{\beta}}$,

$$\begin{aligned}
\mathrm{SSE}_0 - \mathrm{SSE}_1 &= (\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}})^{\mathrm{T}} \left[\boldsymbol{\Lambda}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{\Lambda}^{\mathrm{T}}\right]^{-1} \boldsymbol{\Lambda}\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{\Lambda}^{\mathrm{T}} \left[\boldsymbol{\Lambda}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{\Lambda}^{\mathrm{T}}\right]^{-1} \boldsymbol{\Lambda}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \\
&= \boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y},
\end{aligned}$$

[8] See Christensen (2011, Section 3.3) or Monahan (2008, Section 6.3–6.5).

where $\boldsymbol{A} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{\Lambda}^{\mathrm{T}}$ (Doss 2010). The projection matrix $(\boldsymbol{P}_1 - \boldsymbol{P}_0)$ is $\boldsymbol{A}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}$ for $\boldsymbol{A}$ as just defined.

A yet more general form of the general linear hypothesis is H_0: $\boldsymbol{\Lambda\beta} = \boldsymbol{c}$ for constants $\boldsymbol{c}$. In the F test statistic, we then merely replace $(\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}} - \boldsymbol{0})$ by $(\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}} - \boldsymbol{c})$. This more general H_0 is useful for inverting significance tests to construct confidence regions (Exercise 3.18). Another useful application is *noninferiority testing* in drug research, which analyzes whether the effect of a new drug falls within some acceptable margin c of the effect for an established drug.

3.2.8 Example: Testing That a Single Model Parameter Equals Zero

A common inference in linear modeling is testing H_0: $\beta_j = 0$ that a single explanatory variable in the model can be dropped. This is the special case of H_0: $\boldsymbol{\Lambda\beta} = \boldsymbol{0}$ that substitutes for $\boldsymbol{\Lambda}$ a row vector $\boldsymbol{\lambda}$ with a multiple 1 of β_j and 0 elsewhere. Since the denominator of the F test statistic for comparing two nested models is s^2 (the error mean square) for the full model, the F test statistic then simplifies to

$$F = \frac{(\mathrm{SSE}_0 - \mathrm{SSE}_1)/1}{\mathrm{SSE}_1/(n-p)} = \frac{(\boldsymbol{\lambda}\hat{\boldsymbol{\beta}})^{\mathrm{T}}\left[\boldsymbol{\lambda}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{\lambda}^{\mathrm{T}}\right]^{-1}\boldsymbol{\lambda}\hat{\boldsymbol{\beta}}}{s^2} = \frac{\hat{\beta}_j^2}{(SE_j)^2},$$

where SE_j denotes the standard error of $\hat{\beta}_j$, the square of which is s^2 times the element from the corresponding row and column of $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$. This test statistic has $df_1 = 1$ and $df_2 = n - p$.

In the first ratio in this expression, $(\mathrm{SSE}_0 - \mathrm{SSE}_1)$ is the partial sum of squares explained by adding term j to the linear predictor, once the other terms are already there. The last ratio is $F = t^2$, where $t = \hat{\beta}_j/(SE_j)$. The null distribution of this t statistic is the t distribution with $df = n - p$.

3.2.9 Testing Terms in an Unbalanced Factorial ANOVA

In Section 3.2.1 (Table 3.1) we showed sum-of-squares formulas for the sources in the one-way layout. Analogous relatively simple formulas occur in factorial ANOVA with two or more factors, in the balanced case of equal sample sizes in the cells (e.g., Exercise 3.13). Unbalanced cases do not yield such formulas.

Consider, for example, the two-way layout in which y_{ijk} is observation k in the cell for level i of factor A and level j of factor B, for $i = 1, \ldots, r, j = 1, \ldots, c, k = 1, \ldots, n_{ij}$, where n_{ij} varies with i and j. The model with linear predictor

$$E(y_{ijk}) = \beta_0 + \beta_i + \gamma_j + \delta_{ij}$$

permits interaction between A and B in their effects on y. To achieve identifiability, we can express this as a linear model in which $r - 1$ of $\{\beta_i\}$ are coefficients of indicator variables for all except one level of A, $c - 1$ of $\{\gamma_j\}$ are coefficients of indicator

variables for all except one level of B, and $(r-1)(c-1)$ of $\{\delta_{ij}\}$ are coefficients of products of the $r-1$ indicator variables for A with the $c-1$ indicator variables for B. With unbalanced data, a simple formula no longer occurs for the partial sum of squares explained by the interaction terms, or when those terms are not in the model, by the main effects. However, it is straightforward to fit the full model, fit a reduced model such as with $\{\delta_{ij}=0\}$, and then conduct the F test to compare these two nested models.

More complex models have several factors as well as higher-order interactions. Moreover, some combinations of the factors may have no observations, or the levels of some factors may be nested in levels of other factors, and the model may also contain quantitative explanatory variables. It may not even be obvious how to constrain parameters to achieve identifiability. Good software properly determines this, when we enter the terms as predictors in the linear model. Then we can test whether we need high-order terms in the model by fitting the model with and without those terms and using the F test for nested models to evaluate whether the partial SS explained by those terms is statistically significant. That test is a very general and useful one.

3.3 CONFIDENCE INTERVALS AND PREDICTION INTERVALS FOR NORMAL LINEAR MODELS

We learn more from constructing confidence intervals for parameter values than from significance testing. A confidence interval shows us the entire range of plausible values for a parameter, rather than focusing merely on whether a particular value is plausible.

3.3.1 Confidence Interval for a Parameter of a Normal Linear Model

To construct a confidence interval for a parameter β_j in a normal linear model, we construct and then invert a t test of H_0: $\beta_j = \beta_{j0}$ about potential values for β_j. The test statistic is

$$t = \frac{\hat{\beta}_j - \beta_{j0}}{SE_j},$$

the number of standard errors that $\hat{\beta}_j$ falls from β_{j0}. Recall that SE_j is the square root of the element in row j and column j of the estimated covariance matrix $s^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ of $\hat{\boldsymbol{\beta}}$, where s^2 is the error mean square. Just as the residuals are orthogonal to the model space, the residuals are uncorrelated with $\hat{\boldsymbol{\beta}}$. Specifically, the $p \times n$ covariance matrix

$$\mathrm{cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{y} - \hat{\boldsymbol{\mu}}) = \mathrm{cov}\left[(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}, (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}\right] = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\sigma^2\boldsymbol{I}(\boldsymbol{I}-\boldsymbol{H})^{\mathrm{T}},$$

and this is $\boldsymbol{0}$ because $\boldsymbol{HX} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{X}$. Being linear functions of $\boldsymbol{y}$, $\hat{\boldsymbol{\beta}}$ and $(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$ are jointly normally distributed, so uncorrelatedness implies independence.

Since s^2 is a function of the residuals, $\hat{\boldsymbol{\beta}}$ and s^2 are independent, and so are the numerator and denominator of the t statistic, as is required to obtain a t distribution.

The $100(1-\alpha)\%$ confidence interval for β_j is the set of all β_{j0} values for which the test has P-value $> \alpha$, that is, for which $|t| < t_{\alpha/2,n-p}$, the $1-\alpha/2$ quantile of the t distribution having $df = n-p$. For example, the 95% confidence interval is

$$\hat{\beta}_j \pm t_{0.025,n-p}(SE_j).$$

3.3.2 Confidence Interval for $E(y) = \boldsymbol{x}_0\boldsymbol{\beta}$

At a fixed setting $\boldsymbol{x}_0$ (a row vector) for the explanatory variables, we can construct a confidence interval for $E(y) = \boldsymbol{x}_0\boldsymbol{\beta}$. We do this by constructing and then inverting a t test about values for that linear predictor.

Let $\hat{\mu} = \boldsymbol{x}_0\hat{\boldsymbol{\beta}}$. Now

$$\text{var}(\hat{\mu}) = \text{var}(\boldsymbol{x}_0\hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0\text{var}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_0^{\text{T}} = \sigma^2\boldsymbol{x}_0(\boldsymbol{X}^{\text{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_0^{\text{T}}.$$

Since $\boldsymbol{x}_0\hat{\boldsymbol{\beta}}$ is a linear function of $\boldsymbol{y}$, it has a normal distribution. Thus,

$$z = \frac{\boldsymbol{x}_0\hat{\boldsymbol{\beta}} - \boldsymbol{x}_0\boldsymbol{\beta}}{\sigma\sqrt{\boldsymbol{x}_0(\boldsymbol{X}^{\text{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_0^{\text{T}}}} \sim N(0,1),$$

and

$$t = \frac{\boldsymbol{x}_0\hat{\boldsymbol{\beta}} - \boldsymbol{x}_0\boldsymbol{\beta}}{s\sqrt{\boldsymbol{x}_0(\boldsymbol{X}^{\text{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_0^{\text{T}}}} = \frac{\boldsymbol{x}_0\hat{\boldsymbol{\beta}} - \boldsymbol{x}_0\boldsymbol{\beta}}{\sigma\sqrt{\boldsymbol{x}_0(\boldsymbol{X}^{\text{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_0^{\text{T}}}} \Big/ \sqrt{\frac{s^2}{\sigma^2}} \sim t_{n-p}.$$

This last result follows because $(n-p)s^2/\sigma^2$ has a χ^2_{n-p} distribution for a normal linear model, by Cochran's theorem, so the t statistic is a $N(0,1)$ variate divided by the square root of the ratio of a χ^2_{n-p} variate to its df value. Also, since s^2 and $\hat{\boldsymbol{\beta}}$ are independent, so are the numerator and denominator of the t statistic. It follows that a $100(1-\alpha)\%$ confidence interval for $E(y) = \boldsymbol{x}_0\boldsymbol{\beta}$ is

$$\boldsymbol{x}_0\hat{\boldsymbol{\beta}} \pm t_{\alpha/2,n-p}s\sqrt{\boldsymbol{x}_0(\boldsymbol{X}^{\text{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_0^{\text{T}}}. \tag{3.2}$$

When $\boldsymbol{x}_0$ is the explanatory variable value $\boldsymbol{x}_i$ for a particular observation, the term under the square root is the leverage h_{ii} from the model's hat matrix.

The construction for this interval extends directly to confidence intervals for linear combinations $\boldsymbol{\ell}'\boldsymbol{\beta}$. An example is a contrast of the parameters, such as $\beta_j - \beta_k$ for a pair of levels of a factor.

3.3.3 Prediction Interval for a Future y

At a particular value $\boldsymbol{x}_0$, how can we form an interval that is very likely to contain a future observation y at that value? This is more challenging than forming a confidence

interval for the expected response. With lots of data, we can make precise inference about the mean but not precise prediction about a single future observation.

The normal linear model states that a future value y satisfies

$$y = \boldsymbol{x}_0\boldsymbol{\beta} + \epsilon, \quad \text{where} \quad \epsilon \sim N(0, \sigma^2).$$

From the fit of the model, the prediction of the future y value is $\hat{\mu} = \boldsymbol{x}_0\hat{\boldsymbol{\beta}}$. Now the future y also satisfies

$$y = \boldsymbol{x}_0\hat{\boldsymbol{\beta}} + e, \quad \text{where} \quad e = y - \hat{\mu}$$

is the residual for that observation. Since the future y is independent of the observations $y_1, \ldots, y_n$ used to determine $\hat{\boldsymbol{\beta}}$ and then $\hat{\mu}$,

$$\text{var}(e) = \text{var}(y - \hat{\mu}) = \text{var}(y) + \text{var}(\hat{\mu}) = \sigma^2[1 + \boldsymbol{x}_0(\boldsymbol{X}^{\text{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_0^{\text{T}}].$$

It follows that

$$\frac{y - \hat{\mu}}{\sigma\sqrt{1 + \boldsymbol{x}_0(\boldsymbol{X}^{\text{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_0^{\text{T}}}} \sim N(0, 1) \quad \text{and} \quad \frac{y - \hat{\mu}}{s\sqrt{1 + \boldsymbol{x}_0(\boldsymbol{X}^{\text{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_0^{\text{T}}}} \sim t_{n-p}.$$

Inverting this yields a $100(1 - \alpha)\%$ *prediction interval* for the future y observation,

$$\hat{\mu} \pm t_{\alpha/2,n-p}s\sqrt{1 + \boldsymbol{x}_0(\boldsymbol{X}^{\text{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_0^{\text{T}}}. \tag{3.3}$$

3.3.4 Example: Confidence Interval and Prediction Interval for Simple Linear Regression

We illustrate the confidence interval for the mean and the prediction interval for a future observation with the bivariate linear model,

$$E(y_i) = \beta_0 + \beta_1 x_i.$$

It is simpler to use the explanatory variable in centered form $x_i^* = x_i - \bar{x}$, which (from Section 2.1.3) results in uncorrelated $\hat{\beta}_0$ and $\hat{\beta}_1$. For the centered predictor values, $\hat{\beta}_0$ changes value to $\bar{y}$, but $\hat{\beta}_1$ and $\text{var}(\hat{\beta}_1) = \sigma^2/[\sum_i (x_i - \bar{x})^2]$ do not change. So, at a particular value x_0 for x,

$$\begin{aligned}\text{var}(\hat{\mu}) &= \text{var}[\hat{\beta}_0 + \hat{\beta}_1(x_0 - \bar{x})] \\ &= \text{var}(\bar{y}) + (x_0 - \bar{x})^2\text{var}(\hat{\beta}_1) = \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right].\end{aligned}$$

For a future observation y and its independent prediction $\hat{\mu}$,

$$\text{var}(y - \hat{\mu}) = \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right].$$

The variances are smallest at $x_0 = \bar{x}$ and increase in a symmetric quadratic manner as x_0 moves away from $\bar{x}$. At $x_0 = \bar{x}$, we see that $\text{var}(\hat{\mu}) = \text{var}(\bar{y}) = \sigma^2/n$, whereas $\text{var}(y - \hat{\mu}) = \sigma^2(1 + 1/n)$. As n increases, $\text{var}(\hat{\mu})$ decreases toward 0, but $\text{var}(y - \hat{\mu})$ has σ^2 as its lower bound. Even if we can estimate nearly perfectly the regression line, we are limited in how accurately we can predict any future observation.

Figure 3.1 sketches the confidence interval and prediction interval, as a function of x_0. As n increases, the width of a confidence interval for the mean at any x_0 decreases toward 0, but the width of the 95% prediction interval decreases toward $2(1.96)\sigma$.

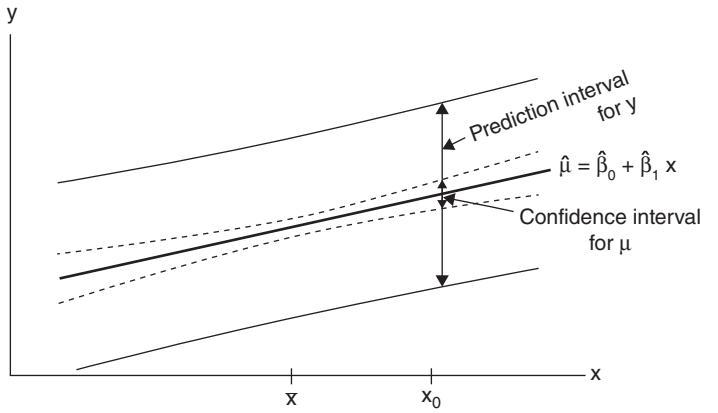

Figure 3.1 Portrayal of confidence intervals for the mean, $E(y) = \beta_0 + \beta_1 x_0$, and prediction intervals for a future observation y, at various x_0 values.

3.3.5 Interpretation and Limitations of Prediction Intervals

Interpreting a prediction interval is awkward. With $\alpha = 0.05$, we would like to say that conditional on the observed data and the model fit, we have 95% confidence that the future y will fall in the interval; that is, close to 95% of a large number of future observations would fall in the interval. However, the probability distributions in the derivation of Section 3.3.3 treat $\hat{\mu}$ as well as the future y as random, whereas in practice we use the interval after observing the data and hence $\hat{\mu}$. The conditional probability that the prediction interval captures a future y, given $\hat{\mu}$, is not 0.95. From the reasoning that led to Equation 3.3, before collecting any data, for the $\hat{\mu}$ (and s) to be found and then the future y,

$$P\left[|y - \hat{\mu}|/s\sqrt{1 + \boldsymbol{x}_0(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_0^{\mathrm{T}}} \le t_{0.025,n-p}\right] = 0.95.$$

Once we observe the data and find $\hat{\mu}$ and s, this probability (with y as the only random part) does not equal 0.95. It depends on where $\hat{\mu}$ happened to fall. It need not be

close to 0.95 unless var($\hat{\mu}$) is negligible compared to var(y). The 95% confidence for a prediction interval means the following: If we repeatedly used this method with many such datasets of independent observations satisfying the model (i.e., to construct both the fitted equation and this interval) and each time made a future observation, in the long run 95% of the time the interval formed would contain the future observation.

To this interpretation, we add the vital qualifier, *if the model truly holds*. In practice, we should have considerable faith in the model before forming prediction intervals. Even if we do not truly believe the model (the usual situation in practice), a confidence interval for $E(y) = \boldsymbol{x}_0\boldsymbol{\beta}$ at various $\boldsymbol{x}_0$ values is useful for describing the fit of the model in the population of interest. However, if the model fails, either in its description of the population mean as a function of the explanatory variables or in its assumptions of normality with constant variance, then the actual percentage of many future observations that fall within the limits of 95% prediction intervals may be quite different from 95%.

3.4 EXAMPLE: NORMAL LINEAR MODEL INFERENCE

What affects the selling price of a house? Table 3.3 shows observations on recent home sales in Gainesville, Florida. This table shows data for 8 houses from a data file for 100 home sales at the text website. Variables listed are selling price (in thousands of dollars), size of house (in square feet), annual property tax bill (in dollars), number of bedrooms, number of bathrooms, and whether the house is new. Since these 100 observations are from one city alone, we cannot use them to make inferences about the relationships in general. But for illustrative purposes, we treat them as a random sample of a conceptual population of home sales in this market and analyze how selling price seems to relate to these characteristics. We suggest that you download the data from the text website, so you can construct graphics not shown here and fit various models that seem sensible.

Table 3.3 Selling Prices and Related Characteristics for a Sample of Home Sales in Gainesville, Florida

Home	Selling Price	Size	Taxes	Bedrooms	Bathrooms	New
1	279.9	2048	3104	4	2	No
2	146.5	912	1173	2	1	No
3	237.7	1654	3076	4	2	No
4	200.0	2068	1608	3	2	No
5	159.9	1477	1454	3	3	No
6	499.9	3153	2997	3	2	Yes
7	265.5	1355	4054	3	2	No
8	289.9	2075	3002	3	2	Yes

Complete file for 100 homes is file `Houses.dat` at `www.stat.ufl.edu/~aa/glm/data`.

3.4.1 Inference for Modeling House Selling Prices

For modeling, we take y = selling price. Section 4.6 discusses issues in selecting explanatory variables for a model. For now, for simplicity we use only x_1 = size of house and x_2 = whether the house is new (1 = yes, 0 = no). We refer to these as "size" and "new." To begin, let us look at the data.

```
------------------------------------------------------------------------
> Houses # complete data at www.stat.ufl.edu/~aa/glm/data
   case  taxes  beds  baths  new  price  size
1     1   3104     4      2    0  279.9  2048
2     2   1173     2      1    0  146.5   912
...
> cbind(mean(price), sd(price), mean(size), sd(size))
            [,1]       [,2]          [,3]      [,4]
[1,]      155.33     101.26       1629.28    666.94
> table(new)
new
 0  1
89 11
> pch.list <- rep(0, 100)
> pch.list[new=="0"] <- 1; pch.list[new=="1"] <- 4 # pick symbols
> plot(size, price, pch=(pch.list)) # plot with symbols for new=0,1
------------------------------------------------------------------------
```

Figure 3.2 shows roughly an increasing linear trend for selling price as a function of size. An exception is a relatively low selling price for a very large dwelling that was not new (observation 64 in the data file). Only 11 houses in the sample were new, so the impact of that variable is rather unclear.

We next fit the model $E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$, having additive effects of these explanatory variables. The least squares fit is $\hat{\mu}_i = -40.231 + 0.116x_{i1} + 57.736x_{i2}$.

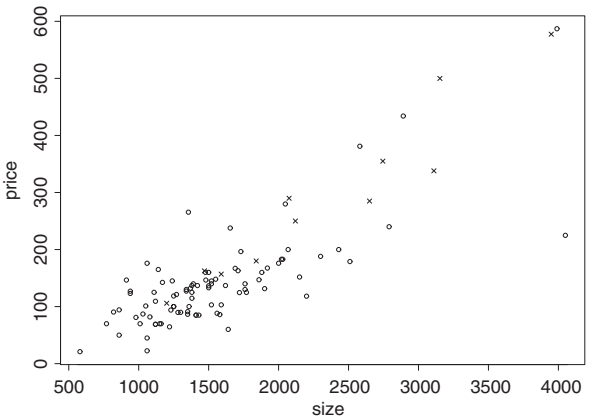

Figure 3.2 Scatterplot of selling price (in thousands of dollars) versus size of house (in square feet), labeled by whether new (× symbol for "yes" and o symbol for "no").

Adjusting for house size, the estimated mean selling price is \$57,736 higher for new homes. Because only 11 houses in the sample were new, this estimate is imprecise. For new or older houses, the estimated mean selling price increases by \$116 for each additional square foot of size. The sample R^2 value is large (0.72).

```
------------------------------------------------------------------------
> fit <- lm(price ~ size + new)
> summary(fit)
Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   -40.2309      14.6961    -2.738    0.00737
size            0.1161       0.0088    13.204    < 2e-16
new            57.7363      18.6530     3.095    0.00257
---
Residual standard error: 53.88 on 97 degrees of freedom # This is s
Multiple R-squared: 0.7226, Adjusted R-squared: 0.7169
F-statistic: 126.3 on 2 and 97 DF, p-value: < 2.2e-16
> plot(fit)
------------------------------------------------------------------------
```

Consider H_0: $\beta_1 = \beta_2 = 0$, stating that neither size nor new has an effect on selling price. The global F test statistic equals 126.3, with $df_1 = 2$ (since there are two effect parameters) and $df_2 = 100 - 3 = 97$. The P-value is 0 to many decimal places. This is no surprise. With this global test, H_0 states that *none* of the explanatory variables are truly correlated with the response. We usually expect a small P-value, and of greater interest is whether each explanatory variable has an effect, adjusting for the other explanatory variables in the model. The t statistic for testing the effect of whether the house is new, adjusting for size, is $t = 3.095$ ($df = 97$), highly significant ($P = 0.003$). Likewise, size has a highly significant partial effect, which again is no surprise.

Next we find a 95% confidence interval for the mean selling price of new homes at the mean size of the new homes, 2354.73 square feet. If the model truly holds, Equation 3.2 implies 95% confidence that the conceptual population mean selling price falls between \$258,721 and \$323,207. Equation 3.3 predicts that a selling price for another new house of that size will fall between \$179,270 and \$402,658.

```
------------------------------------------------------------------------
> predict(fit,data.frame(size=2354.73, new=1), interval="confidence")
      fit        lwr      upr # 95% confidence is default
1 290.964   258.7207 323.2072
> predict(fit,data.frame(size=2354.73, new=1), interval="prediction")
      fit        lwr      upr
1 290.964   179.2701 402.6579
------------------------------------------------------------------------
```

3.4.2 Model Checking

We next check the adequacy of the normal linear model and highlight influential observations. When the model holds, the standardized residuals have approximately a $N(0, 1)$ distribution. Let us look at a histogram and a *Q–Q plot*. The latter plots the

standardized residual values against expected values of order statistics from a $N(0, 1)$ distribution (so-called *normal scores*). When a normal linear model holds, the points should lie roughly on a line through the origin with slope 1. Severe departures from that line indicate substantial non-normality in the conditional distribution of y. However, be cautious in interpreting such plots when n is not large, as they are affected by ordinary sampling variability.

```
------------------------------------------------------------------------
> hist(rstandard(fit)) # use rstudent instead for Studentized residuals
> qqnorm(rstandard(fit)) # Q-Q plot of standardized residuals
------------------------------------------------------------------------
```

For these data, the histogram in Figure 3.3 suggests that the conditional distribution of y is mound shaped, but possibly skewed to the right. Also, observation 64 has a relatively large negative standardized residual of -4.2. The Q–Q plot also shows evidence of skew to the right, because large positive theoretical quantiles have sample quantiles that are larger in absolute value whereas large negative theoretical quantiles have sample quantiles that are smaller in absolute value (except for the outlier). However, it is difficult to judge shape well unless n is quite large, and the actual error rate for two-sided statistical inference about β_j parameters in the linear model is *robust* to violations of the normality assumption. Inadequacy of statistical inference and consequent substantive conclusions are usually affected more by an inappropriate linear predictor (e.g., lacking an important interaction) and by practical sampling problems (e.g., missing data, errors of measurement) than by non-normality of the response. With clearly non-normal residuals, one can transform y to improve the normality. But the linear predictor may then more poorly describe the relationship, and effects on $E[g(y)]$ are of less interest than effects on $E(y)$. So, we recommend

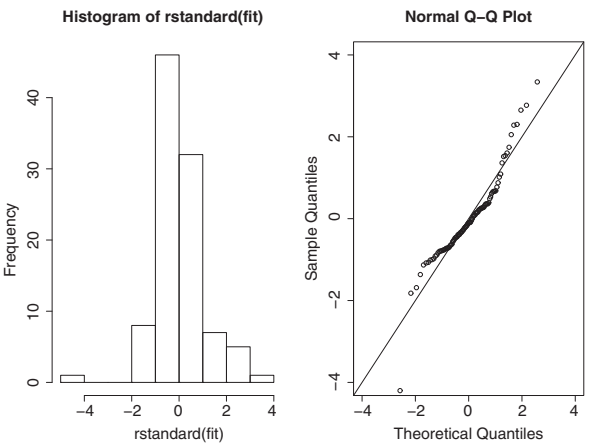

Figure 3.3 Histogram and Q–Q plot of standardized residuals, for normal linear model predicting selling price using size and new as explanatory variables.

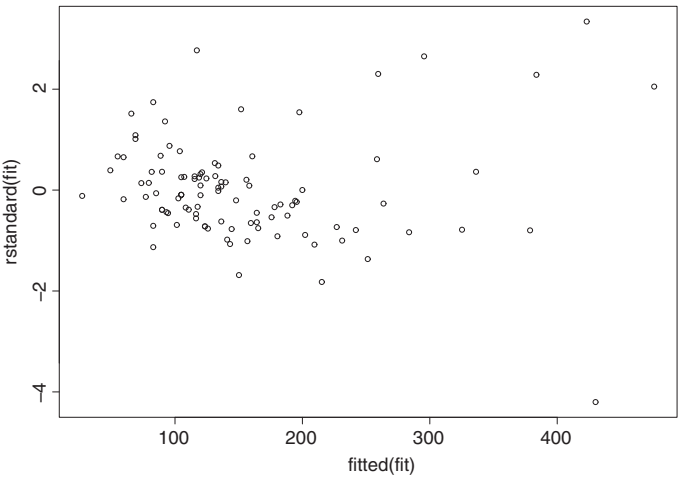

Figure 3.4 Plot of standardized residuals versus fitted values, for linear model predicting selling price using size and new as explanatory variables.

such plots mainly to help detect unusual observations that could influence substantive conclusions.

To investigate the adequacy of the linear predictor, we plot the residuals against the fitted values (Figure 3.4) and against size.

```
------------------------------------------------------------------------
> plot(fitted(fit), rstandard(fit))
> plot(size, rstandard(fit))
------------------------------------------------------------------------
```

If the normal linear model holds, a plot of the residuals against fitted values or values of explanatory variables should show a random pattern about 0 with relatively constant variability (Section 2.5.2). Figure 3.4 also highlights the unusual observation 64, but generally does not indicate lack of fit. There is a suggestion that residuals may tend to be larger in absolute value at higher values of the response. Rather than constant variance, it seems plausible that the variance may be larger at higher mean selling prices. We address this when we revisit the data in the next chapter.

The next table shows some standardized residuals and values of Cook's distance, including results for observation 64, which has the only Cook's distance exceeding 1.

```
------------------------------------------------------------------------
> cooks.distance(fit)
> plot(cooks.distance(fit))
> cbind(case,size,new,price,fitted(fit),rstandard(fit),cooks.distance(fit))
    case size new price
1      1 2048   0 279.9 197.607   1.541 1.462e-02
```

```
2      2  912   0 146.5  65.681   1.517 1.703e-02
...
64    64 4050   0 225.0 430.102  -4.202 1.284e+00
...
-----------------------------------------------------------------------
```

To check whether observation 64 is influential, we refit the model without it.

```
-----------------------------------------------------------------------
> fit2 <- lm(price ~ size + new, subset(Houses, case != 64))
> summary(fit2)
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  -63.1545     14.2519   -4.431  2.49e-05
size           0.1328      0.0088   15.138   < 2e-16
new           41.3062     17.3269    2.384    0.0191
---
Residual standard error: 48.99 on 96 degrees of freedom
Multiple R-squared: 0.772,      Adjusted R-squared: 0.7672
-----------------------------------------------------------------------
```

The effect of a house being new has diminished from \$57,736 to \$41,306, the effect of size has increased some, and R^2 has increased considerably. This observation clearly is influential. We will see that it is not influential or even unusual when we consider an alternative model in Section 4.7.3 that allows the variability to grow with the mean.

There is no assurance that the effects of these two explanatory variables are truly additive. Perhaps the effect of size is different for new houses than for others. We can check by adding an interaction term, which we do for the dataset without the highly influential observation 64:

```
-----------------------------------------------------------------------
> fit3 <- lm(price ~ size + new + size:new, subset(Houses, case != 64))
> summary(fit3)
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  -48.2431     15.6864   -3.075   0.00274
size           0.1230      0.0098   12.536   < 2e-16
new          -52.5122     47.6303   -1.102   0.27303
size:new       0.0434      0.0206    2.109   0.03757
---
Residual standard error: 48.13 on 95 degrees of freedom
Multiple R-squared: 0.7822,     Adjusted R-squared: 0.7753
-----------------------------------------------------------------------
```

Adjusted R^2 increases only from 0.767 to 0.775. The SSE values, not reported here (but available in R by requesting `deviance(fit2)` and `deviance(fit3)`), are 230,358 and 220,055. The F test comparing the two models has test statistic $F = t^2 = 2.109^2 = 4.45$ with $df_1 = 1$ and $df_2 = 95$, giving a P-value = 0.038. This

model estimates that the effect of size is 0.123 for older houses and 0.123 + 0.043 = 0.166 for newer houses. The statistically significant improved fit at the 0.05 level must be weighed against a practically insignificant increase in R^2 and a relatively wide confidence interval for the true difference in size effects for new and older houses.

3.4.3 Conditional versus Marginal Effects: Simpson's Paradox

Alternatively, we could continue with the complete dataset of 100 observations and check whether an improved fit occurs from fitting other models. We might expect that the number of bedrooms is an important predictor of selling price, yet it was not included in the above model. Does it help to include "beds" in the model?

```
------------------------------------------------------------------------
> cor(beds, price)
[1] 0.3940
> summary(lm(price ~ beds))
Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   -28.41       44.30   -0.641     0.523
beds           61.25       14.43    4.243  5.01e-05
---
> fit4 <- lm(price ~ size + new + beds)
> summary(fit4)
Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) -25.1998     25.6022   -0.984   0.32745
size          0.1205      0.0107   11.229   < 2e-16
new          54.8996     19.1128    2.872   0.00501
beds         -7.2927     10.1588   -0.718   0.47458
---
Residual standard error: 54.02 on 96 degrees of freedom
Multiple R-squared: 0.7241,     Adjusted R-squared: 0.7155
------------------------------------------------------------------------
```

Although the number of bedrooms has correlation 0.394 with selling price and is highly significant on its own, it has a P-value of 0.47 for its partial effect. Moreover, the adjusted $R^2 = 0.7155$ is smaller than the value 0.7169 without beds in the model. Apparently once size and new are explanatory variables in the model, it does not help to add beds.

Although the marginal effect of beds is positive, as described by the moderate positive correlation, the estimated partial effect of beds is negative! This illustrates *Simpson's paradox*[9]: An effect of a variable can change direction after adjusting for other variables. Figure 3.5 is a simplistic illustration of how this can happen.

[9]The name refers to Simpson (1951), but the result had been shown by Yule (1903).

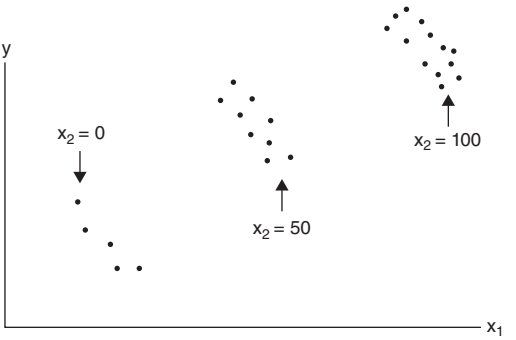

Figure 3.5 Portrayal of Simpson's paradox: The effect of x_1 on y is positive marginally but negative after adjusting for x_2.

3.4.4 Partial Correlation

The partial correlation between selling price and beds while adjusting for size and new is obtained by (1) finding the residuals for predicting selling price using size and new, (2) finding the residuals for predicting beds using size and new, and then (3) finding the ordinary correlation between these two sets of residuals:

```
-----------------------------------------------------------------------
> cor(resid(lm(price ~ size + new)), resid(lm(beds ~ size + new)))
[1] -0.07307201 # partial correlation between selling price and beds
> summary(lm(resid(lm(price ~ size+new)) ~ resid(lm(beds ~ size+new))))
Coefficients: # this yields partial effect of beds on selling price
                              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)                  1.019e-14   5.346e+00    0.000      1.00
resid(lm(beds ~ size + new)) -7.293e+00   1.005e+01   -0.725      0.47
> plot(resid(lm(beds ~ size + new)), resid(lm(price ~ size + new)))
-----------------------------------------------------------------------
```

The partial correlation value of -0.073 is weak. When a true partial correlation is 0, the standard error of a sample partial correlation r for a normal linear model with p parameters is $\sqrt{(1-r^2)/(n-p)}$, about 0.1 in this case.

Using the fact that the multiple correlation $R = \text{corr}(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$, we use the formula at the end of Section 2.5.7 to find the squared partial correlation:

```
-----------------------------------------------------------------------
> (cor(price,fitted(fit4))^2 - cor(price,fitted(fit))^2)/
+                         (1 - cor(price,fitted(fit))^2)
[1] 0.005339518
-----------------------------------------------------------------------
```

The proportion of the variation in selling price unexplained by size and new that is explained by adding beds to the model is only $(-0.073)^2 = 0.0053$. Again, you can

check that effects change substantially if you refit the model without observation 64 (e.g., the partial correlation changes to −0.240).

3.4.5 Testing Contrasts as a General Linear Hypothesis

For a factor in a model, we can test whether particular parameters are equal by expressing the null hypothesis as a set of contrasts. Such a hypothesis has the form of the general linear hypothesis H_0: $\mathbf{\Lambda\beta} = \mathbf{0}$. To illustrate, the analysis that suggested a lack of effect for beds, adjusting for size and new, investigated the linear effect. We could instead treat beds as a factor, with levels (2,3,4,5), to allow a nonlinear impact. Testing whether 3, 4, and 5 bedrooms have the same effect has a null hypothesis consisting of two contrasts and yields a F statistic with $df_1 = 2$ and $df_2 = 94$. The following code shows the contrasts expressed by equating the parameters for 3 bedrooms and 5 bedrooms and equating the parameters for 4 bedrooms and 5 bedrooms, for R constraints that set the parameter for 2 bedrooms equal to 0.

```
------------------------------------------------------------------------
> fit5 <-lm(price ~ size + new + factor(beds))
> Lambda <- matrix(c(0,0,0,0,0,1,0,0,1,-1,-1), nrow=2)
> Lambda
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0    0    0    1    0   -1 # betas for intercept, size, new,
[2,]    0    0    0    0    1   -1 #          beds=3, beds=4, beds=5
> library(car)
> linearHypothesis(fit5, Lambda, test=c("F"))
Hypothesis:
factor(beds)3 - factor(beds)5 = 0
factor(beds)4 - factor(beds)5 = 0
  Res.Df     RSS  Df  Sum of Sq       F  Pr(>F)
1     96  275849
2     94  273722   2     2127.4  0.3653   0.695
------------------------------------------------------------------------
```

3.4.6 Selecting or Building a Model

This chapter has presented inferences for normal linear models but has not discussed how to select a model or build a model from a set of potential explanatory variables. These issues are relevant for all generalized linear models (GLMs), and we discuss them in the next chapter (Section 4.6).

3.5 MULTIPLE COMPARISONS: BONFERRONI, TUKEY, AND FDR METHODS

Using a model to compare many groups or to evaluate the significance of many potential explanatory variables in a model can entail a very large number of inferences. For example, in a one-way layout, comparing each pair of c groups involves $c(c-1)/2$

inferences, which is considerable when c itself is large. Even if each inference has a small error probability, the probability may be substantial that at least one inference is in error. In such cases, we can construct the inferences so that the error probability applies to the entire family of inferences rather than to each individual one. For example, in constructing confidence intervals for pairwise comparisons of means, we can provide 95% *family-wise* confidence that the entire set of intervals simultaneously contains the true differences.

3.5.1 Bonferroni Method for Multiple Inferences

A popular way to conduct multiple inferences while controlling the overall error rate is based on a simple inequality shown by the British mathematician George Boole (1854), in an impressive treatise of which several chapters presented laws of probability.

Boole's inequality: Let $E_1, E_2, \ldots, E_t$ be t events in a sample space. Then, the probability that at least one of these events occurs has the upper bound

$$P(\cup_j E_j) \leq \sum_{j=1}^{t} P(E_j).$$

The proof of this is simple. We suggest that you construct a Venn diagram to illustrate. Let

$$B_1 = E_1,\ B_2 = E_1^c \cap E_2,\ B_3 = E_1^c \cap E_2^c \cap E_3, \ldots.$$

Then, $\cup_j B_j = \cup_j E_j$ and $B_j \subset E_j$, but the $\{B_j\}$ are disjoint and so $P(\cup_j B_j) = \sum_j P(B_j)$. Thus,

$$P(\cup_j E_j) = P(\cup_j B_j) = \sum_{j=1}^{t} P(B_j) \leq \sum_{j=1}^{t} P(E_j).$$

In the context of multiple confidence intervals, let E_j (for $j = 1, \ldots, t$) denote the event that interval j is in error, not containing the relevant parameter value. If each interval has confidence coefficient $(1 - \alpha/t)$, then the (a priori) probability that at least one of the t intervals is in error is bounded above by $t(\alpha/t) = \alpha$. So, the family-wise confidence coefficient for the set of the t intervals is bounded below by $1 - \alpha$. For example, for the one-way layout with $c = 5$ means, if we use confidence level 99% for each of the 10 pairwise comparisons, the overall confidence level is at least 90%. This method for constructing simultaneous confidence intervals is called the *Bonferroni method*. It relies merely on Boole's inequality, but the name refers to the Italian probabilist/mathematician Carlo Bonferroni, who in 1936 extended Boole's inequality in various ways.

An advantage of the Bonferroni method is its generality. It applies for any probability-based inferences for any distribution, not just confidence intervals for

a normal linear model. A disadvantage is that the method is *conservative*: If we want overall 90% confidence (say), the method ensures that the actual confidence level is *at least* that high. As a consequence, the intervals are wider than ones that would produce *exactly* that confidence level. The next method discussed is more limited, being designed specifically for comparing means in balanced normal linear models, but it does not have this disadvantage.

3.5.2 Tukey Method of Multiple Comparisons

In 1953 the great statistician John Tukey proposed a method for simultaneously comparing means of several normal distributions. Using a probability distribution for the range of observations from a normal distribution, it applies to balanced designs such as one-way and two-way layouts with equal sample sizes.

Definition. Suppose $\{y_i\}$ are independent, with $y_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, c$. Let s^2 be an independent estimate of σ^2 with $\nu s^2/\sigma^2 \sim \chi^2_\nu$. Then,

$$Q = \frac{max_i\, y_i - min_i\, y_i}{s}$$

has the *Studentized range distribution* with parameters c and ν. We denote the distribution by $Q_{c,\nu}$ and its $1 - \alpha$ quantile by $Q_{1-\alpha,c,\nu}$.

To illustrate how Tukey's method uses the Studentized range distribution, we consider the balanced one-way layout for the normal linear model. The sample means $\bar{y}_1, \ldots, \bar{y}_c$ each have sample size $n_i = n$. Let $N = \sum_i n_i = cn$. Let $s^2 = \sum_{i=1}^{c} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2/(N - c)$ denote the pooled variance estimate from the one-way ANOVA (i.e., the error mean square in Section 3.2.1). Then each $\sqrt{n}(\bar{y}_i - \mu_i)$ has a $N(0, \sigma^2)$ distribution, and so

$$\sqrt{n}[\max_i(\bar{y}_i - \mu_i) - \min_i(\bar{y}_i - \mu_i)]/s \sim Q_{c,N-c}.$$

A priori, the probability is $(1 - \alpha)$ that this statistic is less than $Q_{1-\alpha,c,N-c}$. When the statistic *is* bounded above by $Q_{1-\alpha,c,N-c}$, then

$$\text{all } |(\bar{y}_i - \mu_i) - (\bar{y}_j - \mu_j)| < Q_{1-\alpha,c,N-c}(s/\sqrt{n})$$

and thus $(\mu_i - \mu_j)$ falls within $Q_{1-\alpha,c,N-c}(s/\sqrt{n})$ of $(\bar{y}_i - \bar{y}_j)$ for *all* pairs. So, we can construct family-wise confidence intervals for the pairs $\{\mu_i - \mu_j\}$ using simultaneously for all i and j,

$$(\bar{y}_i - \bar{y}_j) \pm Q_{1-\alpha,c,N-c}\left(\frac{s}{\sqrt{n}}\right).$$

The confidence coefficient for the family of all $t = c(c-1)/2$ such comparisons equals $1-\alpha$. A difference $|\bar{y}_i - \bar{y}_j|$ that exceeds $Q_{1-\alpha,c,N-c}(s/\sqrt{n})$ is considered statistically significant, as the interval for $(\mu_i - \mu_j)$ does not contain 0. The corresponding margin of error using the Bonferroni method is $t_{\alpha/c(c-1),N-c}s\sqrt{2/n}$.

To illustrate, suppose we plan to construct family-wise 95% confidence intervals for the 45 pairs of means for $c = 10$ groups, and we have $n = 20$ observations from each group and a standard deviation estimate of $s = 15$. The margin of error for each comparison is $Q_{0.95,10,190}(15/\sqrt{20}) = 15.19$ for the Tukey method and $t_{0.05/2(45),190}(15\sqrt{2/20}) = 15.71$ for the Bonferroni method. The Q and t quantiles used here are easily obtained with software:

```
------------------------------------------------------------------------
> qtukey(0.95, 10, 190); qt(1 - 0.05/(2*45), 190)
[1] 4.527912
[1] 3.311379
------------------------------------------------------------------------
```

The Tukey method applies exactly to this balanced case, for which the sample means have equal variances. A generalized version applies in a slightly conservative manner for unbalanced cases (see Note 3.5).

3.5.3 Controlling the False Discovery Rate

As the number of inferences (t) increases in multiple comparison methods designed to have fixed family-wise error rate α, the margin of error for each inference increases. When t is enormous, as in detecting differential expression in thousands of genes, there may be very low power for establishing significance with any individual inference. It can be difficult to discover any effects that truly exist, especially if those effects are weak. But, in the absence of a multiplicity adjustment, most significant results found could be Type I errors, especially when the number of true non-null effects is small. Some multiple inference methods attempt to address this issue. Especially popular are methods that control the *false discovery rate* (FDR). In the context of significance testing, this is the expected proportion of the rejected null hypotheses ("discoveries") that are erroneously rejected (i.e., that are actually true—"false discoveries").

Benjamini and Hochberg (1995) proposed a simple algorithm for ensuring FDR $\leq \alpha$ for a desired α. It applies with t independent[10] tests. Let $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(t)}$ denote the ordered P-values for the t tests. We reject hypotheses $(1), \ldots, (j^*)$, where j^* is the maximum j for which $P_{(j)} \leq j\alpha/t$. The actual FDR for this method is bounded above by α times the proportion of rejected hypotheses that are actually true. This bound is α when the null hypothesis is always true.

Here is intuition for comparing $P_{(j)}$ to $j\alpha/t$ in this method: Suppose t_0 of the t hypotheses tested are actually true. Since P-values based on continuous test statistics

[10]Benjamini and Yekutieli (2001) showed that the method also works with tests that are positively dependent in a certain sense.

have a uniform distribution when H_0 is true, conditional on $P_{(j)}$ being the cutoff for rejection, a priori we expect to reject about $t_0 P_{(j)}$ of the t_0 true hypotheses. Of the j observed tests actually having P-value $\leq P_{(j)}$, this is a proportion of expected false rejections of $t_0 P_{(j)}/j$. In practice t_0 is unknown, but since $t_0 \leq t$, if $tP_{(j)}/j \leq \alpha$ then this ensures $t_0 P_{(j)}/j \leq \alpha$. Therefore, rejecting H_0 whenever $P_{(j)} \leq j\alpha/t$ ensures this.

With this method, the most significant test compares $P_{(1)}$ to α/t and has the same decision as in the ordinary Bonferroni method, but then the other tests have less conservative requirements. When some hypotheses are false, the FDR method tends to reject more of them than the Bonferroni method, which focuses solely on controlling the family-wise error rate. Benjamini and Hochberg illustrated the FDR for a study about myocardial infarction. For the 15 hypotheses tested, the ordered P-values were

$$0.0001, 0.0004, 0.0019, 0.0095, 0.020, 0.028, 0.030,$$
$$0.034, 0.046, 0.32, 0.43, 0.57, 0.65, 0.76, 1.00.$$

With $\alpha = 0.05$, these are compared with $j(0.05)/15$, starting with $j = 15$. The maximum j for which $P_{(j)} \leq j(0.0033)$ is $j = 4$, for which $P_{(4)} = 0.0095 < 4(0.0033)$. So the hypotheses with the four smallest P-values are rejected. By contrast, the Bonferroni approach with family-wise error rate 0.05 compares each P-value to $0.05/15 = 0.0033$ and rejects only three of these hypotheses.

CHAPTER NOTES

Section 3.1: Distribution Theory for Normal Variates

3.1 **Cochran's theorem**: Results on quadratic forms in normal variates were shown by the Scottish statistician William Cochran in 1934 when he was a 24-year old graduate student at the University of Cambridge, studying under the supervision of John Wishart. He left Cambridge without completing his Ph.D. degree to work at Rothamsted Experimental Station, recruited by Frank Yates after R. A. Fisher left to take a professorship at University College, London. In the 1934 article, Cochran showed that if $x_1, \ldots, x_n$ are iid $N(0, 1)$ and $\sum_i x_i^2 = Q_1 + \cdots + Q_k$ for quadratic forms having ranks $r_1, \ldots, r_k$, then $Q_1, \ldots, Q_k$ are independent chi-squared with df values $r_1, \ldots, r_k$ if and only if $r_1 + \cdots + r_k = n$.

3.2 **Independent normal quadratic forms**: The Cochran's theorem implication that $\{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{y}\}$ are independent when $\boldsymbol{P}_i\boldsymbol{P}_j = \boldsymbol{0}$ extends to this result (Searle 1997, Chapter 2): When $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$, $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{y}$ and $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{y}$ are independent if and only if $\boldsymbol{AVB} = \boldsymbol{0}$.

Section 3.2: Significance Tests for Normal Linear Models

3.3 **Fisher and ANOVA**: Application of ANOVA was stimulated by the 1925 publication of R. A. Fisher's classic text, *Statistical Methods for Research Workers*. Later contributions include Scheffé (1959) and Hoaglin et al. (1991).

3.4 **General linear hypothesis**: For further details about tests for the general linear hypothesis and in particular for one-way and two-way layouts, see Lehmann and Romano (2005, Chapter 7) and Scheffé (1959, Chapters 2–4).

Section 3.5: Multiple Comparisons: Bonferroni, Tukey, FDR Methods

3.5 **Boole, Bonferroni, Tukey, Scheffé**: Seneta (1992) surveyed probability inequalities presented by Boole and Bonferroni and related results of Fréchet. For an overview of Tukey's contributions to multiple comparisons, see Benjamini and Braun (2002) and Tukey (1994). With unbalanced data, Kramer (1956) suggested replacing $s/\sqrt{n}$ in the Tukey interval by $s\sqrt{\frac{1}{2}\left[(1/n_a)+(1/n_b)\right]}$ for groups a and b. Hayter (1984) showed this is slightly conservative. For the normal linear model, Scheffé (1959) proposed a method that applies simultaneously to all contrasts of c means. For estimating a contrast $\sum_i a_i\mu_i$ in the one-way layout (possibly unbalanced), it multiplies the usual estimated standard error $s\sqrt{\sum_i (a_i^2/n_i)}$ for $\sum_i a_i\bar{y}_i$ by $\sqrt{(c-1)F_{1-\alpha,c-1,n-c}}$ to obtain the margin of error. For simple differences between means, these are wider than the Tukey intervals, because they apply to a much larger family of contrasts. Hochberg and Tamhane (1987) and Hsu (1996) surveyed multiple comparison methods.

3.6 **False discovery rate**: For surveys of FDR methods and issues in large-scale multiple hypothesis testing, see Benjamini (2010), Dudoit et al. (2003), and Farcomeni (2008).

EXERCISES

3.1 Suppose $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$ with $\boldsymbol{V}$ nonsingular of rank p. Show that $(\boldsymbol{y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) \sim \chi^2_p$ by letting $\boldsymbol{z} = \boldsymbol{V}^{-1/2}(\boldsymbol{y}-\boldsymbol{\mu})$ and finding the distribution of $\boldsymbol{z}$ and $\boldsymbol{z}^{\mathrm{T}}\boldsymbol{z}$.

3.2 If T has a t distribution with $df = p$, then using the construction of t and F random variables, explain why T^2 has the F distribution with $df_1 = 1$ and $df_2 = p$.

3.3 Suppose $z = x + y$ where $z \sim \chi^2_p$ and $x \sim \chi^2_q$. Show how to find the distribution of y.

3.4 Applying the SS decomposition with the projection matrix for the null model (Section 2.3.1), use Cochran's theorem to show that for $y_1, \ldots, y_n$ independent from $N(\mu, \sigma^2)$, $\bar{y}$ and s^2 are independent (Cochran 1934).

3.5 For $y_1, \ldots, y_n$ independent from $N(\mu, \sigma^2)$, apply Cochran's theorem to construct a F test of H_0: $\mu = \mu_0$ against H_1: $\mu \neq \mu_0$ by applying the SS decomposition with the projection matrix for the null model shown in Section 2.3.1 to the adjusted observations $\{y_i - \mu_0\}$. State the null and alternative distributions of the test statistic. Show how to construct an equivalent t test.

3.6 Consider the normal linear model for the one-way layout (Section 3.2.1).

a. Explain why the F statistic used to test H_0: $\mu_1 = \cdots = \mu_c$ has, under H_0, an F distribution.

b. Why is the test is called analysis of *variance* when H_0 deals with *means*? (*Hint*: See Section 3.2.5.)

3.7 A one-way ANOVA uses n_i observations from group i, $i = 1, \ldots, c$.

a. Verify the noncentrality parameter for the scaled between-groups sum of squares.

b. Suppose $c = 3$, with $\mu_1 - \mu_2 = \mu_2 - \mu_3 = \sigma/2$. Evaluate the noncentrality, and use it to find the power of a F test with size $\alpha = 0.05$ for a common sample size n, when (i) $n = 10$, (ii) $n = 30$, (iii) $n = 50$.

c. Now suppose $\mu_1 - \mu_2 = \mu_2 - \mu_3 = \Delta\sigma$. Evaluate the noncentrality when each $n_i = 10$, and use it to find the power of a F test with size $\alpha = 0.05$ when $\Delta = 0, 0.5, 1.0$.

3.8 Based on the formula $s^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ for the estimated var$(\hat{\boldsymbol{\beta}})$, explain why the standard errors of $\{\hat{\beta}_j\}$ tend to decrease as n increases.

3.9 Using principles from this chapter, inferentially compare μ_1 and μ_2 from $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ populations, based on independent random samples of sizes n_1 and n_2.

a. Put the analysis in a normal linear model context, showing a model matrix and explaining how to interpret the model parameters.

b. Find the projection matrix for the model space, and find SSR and SSE.

c. Construct a F test statistic for testing H_0: $\mu_1 = \mu_2$ against H_a: $\mu_1 \neq \mu_2$. Using Cochran's theorem, specify a null distribution for this statistic.

d. Relate the F test statistic in (**c**) to the t statistic for this test,

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where s^2 is the pooled variance estimate from the two samples.

3.10 Refer to the previous exercise. Based on inverting significance tests with nonzero null values, show how to construct a confidence interval for $\mu_1 - \mu_2$.

3.11 Section 2.3.4 considered the projection matrices and ANOVA table for the two-way layout with one observation per cell. For testing each main effect in that model, show how to construct test statistics and explain how to obtain their null distributions, based on theory in this chapter.

3.12 For the balanced two-way $r \times c$ layout with n observations $\{y_{ijk}\}$ in each cell, denote the sample means by $\{\bar{y}_{ij.}\}$ in the cells, $\bar{y}_{i..}$ in level i of A, $\bar{y}_{.j.}$ in level j of B, and $\bar{y}$ overall for all $N = nrc$ observations. Consider the model that assumes a lack of interaction.

a. Construct the ANOVA table, including SS and df values, showing how to construct F statistics for testing the main effects.

b. Show that the expected value of the numerator mean square for the test of the A factor effect is $\sigma^2 + \left(\frac{cn}{r-1}\right)\sum_{i=1}^{r}(\mu_{i..} - \bar{\mu})^2$.

3.13 Refer to the previous exercise. Now consider the model permitting interaction. Table 3.4 shows the resulting ANOVA table.

a. Argue intuitively and in analogy with results for one-way ANOVA that the SS values for factor A, factor B, and residual are as shown in the ANOVA table.

b. Based on the results in (**a**) and what you know about the total of the SS values, show that the SS for interaction is as shown in the ANOVA table.

c. In the ANOVA table, show the df values for each source. Show the mean squares, and show how to construct test statistics for testing no interaction and for testing each main effect. Specify the null distribution for each test statistic.

Table 3.4 ANOVA Table for Normal Linear Model with Two-Way Layout

Source	df	Sum of Squares	Mean Square	F_{obs}
Mean	1	$N\bar{y}^2$		
A (rows)	—	$cn\sum_i(\bar{y}_{i..}-\bar{y})^2$	—	—
B (columns)	—	$rn\sum_j(\bar{y}_{.j.}-\bar{y})^2$	—	—
Interaction	—	$n\sum_i\sum_j(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y})^2$	—	—
Residual (error)	—	$\sum_i\sum_j\sum_k(y_{ijk}-\bar{y}_{ij.})^2$	—	
Total	N	$\sum_{i=1}^r\sum_{j=1}^c\sum_{k=1}^n y_{ijk}^2$		

3.14 **a.** Show that the F statistic in Section 3.2.4 for testing that all effects equal 0 has expression in terms of the R^2 value as

$$F=\frac{R^2/(p-1)}{(1-R^2)/(n-p)}$$

b. Show that the F statistic (3.1) for comparing nested models has expression in terms of the R^2 values for the models as

$$F=\frac{(R_1^2-R_0^2)/(p_1-p_0)}{(1-R_1^2)/(n-p_1)}.$$

3.15 Using the F formula for comparing models in the previous exercise, show that adjusted R^2 being larger for the more complex model is equivalent to $F>1$.

3.16 For the linear model $E(y_{ij})=\beta_0+\beta_i$ for the one-way layout, explain how H_0: $\beta_1=\cdots=\beta_c$ is a special case of the general linear hypothesis.

3.17 For a normal linear model with p parameters and n observations, explain how to test H_0: $\beta_j=\beta_k$ in the context of the (**a**) general linear hypothesis and (**b**) F test comparing two nested linear models.

3.18 Explain how to use the F test for the general linear hypothesis H_0: $\boldsymbol{\Lambda\beta} = \boldsymbol{c}$ to invert a test of H_0: $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ to form a *confidence ellipsoid* for $\boldsymbol{\beta}$. For $p = 2$, describe how this could give you information beyond what you would learn from separate intervals for β_1 and β_2.

3.19 Suppose a one-way layout has ordered levels for the c groups, such as dose levels in a dose–response assessment. The model $E(y_{ij}) = \beta_0 + \beta_i$ treats the groups as a qualitative factor. The model $E(y_{ij}) = \beta_0 + \beta x_i$ has a quantitative predictor that assumes monotone group scores $\{x_i\}$.

- **a.** Explain why the quantitative-predictor model is a special case of the qualitative-predictor model. Given the qualitative-predictor model, show how the null hypothesis that the quantitative-predictor model is adequate is a special case of the general linear hypothesis. Illustrate by showing $\boldsymbol{\Lambda}$ for the case $c = 5$ with $\{x_i = i\}$.
- **b.** Explain how to use an F test to compare the models, specifying the *df* values.
- **c.** Describe an advantage and disadvantage of each way of handling ordered categories.

3.20 Mimicking the derivation in Section 3.3.2, derive a confidence interval for the linear combination $\boldsymbol{\ell\beta}$. Explain how it simplifies for the case $\beta_j - \beta_k$.

3.21 When there are no explanatory variables, show how the confidence interval in Section 3.3.2 simplifies to a confidence interval for the marginal $E(y)$.

3.22 Consider the null model, for simplicity with known σ^2. After estimating $\mu = E(y)$ by $\bar{y}$, you plan to predict a future y from the $N(\mu, \sigma^2)$ distribution. State the formula for a 95% prediction interval for this model. Suppose, unknown to you, $\bar{y} = \mu + z_o\sigma/\sqrt{n}$ for some particular z_o value. Find an expression for the actual probability, conditional on $\bar{y}$, that the prediction interval contains the future y. Explain why this is not equal to 0.95 (e.g., what happens if $z_o = 0$?) but converges to it as $n \to \infty$.

3.23 Based on the expression for a squared partial correlation in Section 3.4.4, show how it relates to a partial SS for the full model and SSE for the model without that predictor.

3.24 For the normal linear model for the $r \times c$ two-way layout with n observations per cell, explain how to use the Tukey method for family-wise comparisons of all pairs of the r row means with confidence level 95%.

3.25 An analyst plans to construct family-wise confidence intervals for normal linear model parameters $\{\beta^{(1)}, \ldots, \beta^{(g)}\}$ in estimating an effect as part of a meta-analysis with g independent studies. Explain why constructing each interval

with confidence level $(1-\alpha)^{1/g}$ provides exactly the family-wise confidence level $(1-\alpha)$. Prove that such intervals are narrower than Bonferroni intervals.

3.26 In the one-way layout with c groups and a fixed common sample size n, consider simultaneous confidence intervals for pairwise comparisons of means, using family-wise error probability $\alpha = 0.05$. Using software such as R, analyze how the ratio of margins of error for the Tukey method to the Bonferroni method behaves as c increases for fixed n and as n increases for fixed c. Show that this ratio converges to 1 as α approaches 0 (i.e., the Bonferroni method is only very slightly conservative when applied with very small α).

3.27 *Selection bias*: Suppose the normal linear model $\mu_i = \beta_0 + \beta_1 x_i$ holds with $\beta_1 > 0$, but the responses are *truncated* and we observe y_i only when $y_i > L$ (or perhaps only when $y_i < L$) for some threshold L.

a. Describe a practical scenario for which this could happen. How would you expect the truncation to affect $\hat{\beta}_1$ and s? Illustrate by sketching a graph. (You could check this with data, such as by fitting the model in Section 3.4.1 only to house sales having $y_i > 150$.)

b. Construct a likelihood function with the conditional distribution of y, to enable consistent estimation of $\boldsymbol{\beta}$. (See Amemiya (1984) for a survey of modeling with truncated or censored data. In R, see the `truncreg` package.)

3.28 In the previous exercise, suppose truncation instead occurs on x. Would you expect this to affect (**a**) $E(\hat{\beta}_1)$? (**b**) inference about β_1? Why?

3.29 Construct a Q–Q plot for the model for the house selling prices that uses size, new, and their interaction as the predictors, and interpret. To get a sense of how such a plot with a finite sample size may differ from its expected pattern when the model holds, randomly generate 100 standard normal variates a few times and form a Q-Q plot each time.

3.30 Suppose the relationship between y = college GPA and x = high school GPA satisfies $y_i \sim N(1.80 + 0.40x_i, 0.30^2)$. Simulate and construct a scatterplot for $n = 1000$ independent observations taken from this model when x_i has a uniform distribution (**a**) over (2.0, 4.0), (**b**) over (3.5, 4.0). In each case, find R^2. How do R^2 and corr$(\boldsymbol{x}, \boldsymbol{y})$ depend on the range sampled for $\{x_i\}$? Use the formula for R^2 to explain why this happens.

3.31 Refer to Exercise 1.21 on a study comparing forced expiratory volume (y = *fev*1 in the data file) for three drugs (x_2), adjusting for a baseline measurement (x_1).

a. Fit the normal linear model using both x_1 and x_2 and their interaction. Interpret model parameter estimates.

b. Test to see whether the interaction terms are needed. Interpret using confidence intervals for parameters in your chosen model.

3.32 For the horseshoe crab dataset `Crabs.dat` at the text website, analyze inferentially the effect of color on the mean number of satellites, treating the data as a random sample from a conceptual population of female crabs. Fit the normal one-way ANOVA model using color as a qualitative factor. Report results of the significance test for the color effect, and interpret. Provide evidence that the inferential assumption of a normal response with constant variance is badly violated. (Section 7.5 considers more appropriate models.)

3.33 Refer to Exercise 2.47 on carapace width of attached male horseshoe crabs. Extend your analysis of that exercise by conducting statistical inference, and interpret.

3.34 Section 3.4.1 used x_1 = size of house and x_2 = whether new to predict y = selling price. Suppose we instead use a GLM, $\log(\mu_i) = \beta_0 + \beta_1 \log(x_{i1}) + \beta_2 x_{i2}$.

a. For this GLM, interpret β_1 and β_2. (*Hint*: Adjusting for the other variable, find multiplicative effects on μ_i of (i) changing x_{i2} from 0 to 1, (ii) increasing x_{i1} by 1%.)

b. Fit the GLM, assuming normality for $\{y_i\}$, and interpret. Compare the predictive power of this model with the linear model of Section 3.4.1 by finding $R = \text{corr}(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ for each model.

c. For this GLM or the corresponding LM for $E[\log(y_i)]$, refit the model without the most influential observation and summarize. Also, determine whether the fit improves significantly by permitting interaction between $\log(x_{i1})$ and x_{i2}.

3.35 For the house selling price data of Section 3.4, when we include size, new, and taxes as explanatory variables, we obtain

```
------------------------------------------------------------------
> summary(lm(price ~ size + new + taxes))
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)  -21.3538     13.3115   -1.604   0.11196
size           0.0617      0.0125    4.937   3.35e-06
new           46.3737     16.4590    2.818   0.00588
taxes          0.0372      0.0067    5.528   2.78e-07
---
Residual standard error: 47.17 on 96 degrees of freedom
Multiple R-squared: 0.7896,   Adjusted R-squared: 0.783
F-statistic: 120.1 on 3 and 96 DF, p-value: < 2.2e-16
> anova(lm(price ~ size + new + taxes)) # sequential SS, size first
Analysis of Variance Table
Response: price
          Df  Sum Sq  Mean Sq  F value     Pr(>F)
```

```
size       1  705729   705729  317.165  < 2.2e-16
new        1   27814    27814   12.500  0.0006283
taxes      1   67995    67995   30.558  2.782e-07
Residuals 96  213611     2225
---------------------------------------------------------------
```

a. Report and interpret results of the global test of the hypothesis that none of the explanatory variables has an effect.

b. Report and interpret significance tests for the individual partial effects, adjusting for the other variables in the model.

c. What is the conceptual difference between the test of the size effect in the coefficients table and in the ANOVA table?

3.36 Using the house selling price data at the text website, describe the predictive power of various models by finding adjusted R^2 when (i) size is the sole predictor, (ii) size and new are main-effect predictors, (iii) size, new, and taxes are main-effect predictors, (iv) case (iii) with the addition of the three two-way interaction terms. Of these four, which is the simplest model that seems adequate? Why?

3.37 For the house selling price data, fit the model with size of home as the sole explanatory variable. Find a 95% confidence interval for $E(y)$ and a 95% prediction interval for y, at the sample mean size. Interpret.

3.38 In a study[11] at Iowa State University, a large field was partitioned into 20 equal-size plots. Each plot was planted with the same amount of seed corn, using a fixed spacing pattern between the seeds. The goal was to study how the yield of corn later harvested from the plots depended on the levels of use of nitrogen-based fertilizer (low = 45 kg per hectare, high = 135 kg per hectare) and manure (low = 84 kg per hectare, high = 168 kg per hectare). The corn yields (in metric tons) for this completely randomized two-factor study are shown in the table:

Fertilizer	Manure	Observations, by Plot				
High	High	13.7	15.8	13.9	16.6	15.5
High	Low	16.4	12.5	14.1	14.4	12.2
Low	High	15.0	15.1	12.0	15.7	12.2
Low	Low	12.4	10.6	13.7	8.7	10.9

a. Conduct a two-way ANOVA, assuming a lack of interaction between fertilizer level and manure level in their effects on crop yield. Report the ANOVA table. Summarize the main effect tests, and interpret the P-values.

[11] Thanks to Dan Nettleton, Iowa State University, for data on which this exercise is based.

b. If yield were instead measured in some other units, such as pounds or tons, then in your ANOVA table, what will change and what will stay the same?

c. Follow up the main-effect tests in (**a**) by forming 95% Bonferroni confidence intervals for the two main-effect comparisons of means. Interpret.

d. Now allow for interaction, and show results of the F test of the hypothesis of a lack of interaction. Interpret.

3.39 Refer to the study for comparing instruction methods mentioned in Exercise 2.45. Write a short report summarizing inference for the model fitted there, interpreting results and attaching edited software output as an appendix.

3.40 For the `Student survey.dat` data file at the text website, model how political ideology relates to number of times per week of newspaper reading and religiosity. Prepare a report, posing a research question, and then summarizing your graphical analyses, models and interpretations, inferences, checks of assumptions, and overall summary of the relationships.

3.41 For the anorexia study of Exercise 1.24, write a report in which you pose a research question and then summarize your analyses, including graphical description, interpretation of a model fit and its inferences, and checks of assumptions.