36. Multisample U-statistics and jointly distributed U-statistics

Lehmann §6.1

In this topic, we generalize the idea of U-statistics in two different directions. First, we consider single U-statistics for situations in which there is more than one sample. Next, we consider the joint asymptotic distribution of two (single-sample) U-statistics.

We begin by generalizing the idea of U-statistics to the case in which we have more than one random sample. Suppose that \( X_{1i}, \ldots, X_{ni} \) is an iid sample from \( F_i \) for all \( 1 \leq i \leq s \). In other words, we have \( s \) random samples, each potentially from a different distribution, and \( n_i \) is the size of the \( i \)th sample. We may define a statistical functional

\[
\theta = E \phi (X_{11}, \ldots, X_{1a_1}; X_{21}, \ldots, X_{2a_2}; \ldots; X_{s1}, \ldots, X_{sa_s}) .
\]  

(132)

Notice that some of the arguments \( s \) merely been the constant \( j \), \( j \ldots, \).

By an argument similar to the one used in the proof of Theorem 35.1, but much more tedious notationally, we may show that

\[
U_N = \frac{1}{\binom{n}{a_1}} \cdots \frac{1}{\binom{n}{a_s}} \sum_{1 \leq i_1 < \cdots < i_{a_1} \leq n_1} \phi (X_{1i_1}, \ldots, X_{1a_1}; \ldots; X_{sr_1}, \ldots, X_{sr_{a_s}}) .
\]  

(133)

Note that \( N \) denotes the total of all the sample sizes: \( N = n_1 + \cdots + n_s \).

As we did in the case of single-sample U-statistics, define for \( 0 \leq j_1 \leq a_1, \ldots, 0 \leq j_s \leq a_s \)

\[
\phi_{j_1 \ldots j_s} (X_{11}, \ldots, X_{1j_1}; \ldots; X_{s1}, \ldots, X_{sj_s}) = E \{ \phi (X_{11}, \ldots, X_{1a_1}; \ldots; X_{s1}, \ldots, X_{sa_s}) \mid X_{11}, \ldots, X_{1j_1}, \ldots, X_{s1}, \ldots, X_{sj_s} \}
\]  

(134)

and

\[
\sigma_{j_1 \ldots j_s}^2 = \operatorname{Var} \phi_{j_1 \ldots j_s} (X_{11}, \ldots, X_{1j_1}; \ldots; X_{s1}, \ldots, X_{sj_s}) .
\]  

(135)

By an argument similar to the one used in the proof of Theorem 35.1, but much more tedious notationally, we can show that

\[
\sigma_{j_1 \ldots j_s}^2 = \operatorname{Cov} \{ \phi (X_{11}, \ldots, X_{1a_1}; \ldots; X_{s1}, \ldots, X_{sa_s}),
\phi (X_{11}, \ldots, X_{1j_1}, X_{1,a_1+1}, \ldots, X_{1,a_1+(a_1-j_1)}; \ldots; X_{s1}, \ldots, X_{sj_s}, X_{s,a_s+1}, \ldots, X_{s,a_s+(a_s-j_s)}) \} .
\]  

(136)

Notice that some of the \( j_i \) may equal 0. This was not true in the single-sample case, since \( \phi_0 \) would have merely been the constant \( \theta \), so \( \sigma_0^2 = 0 \).

In the special case when \( s = 2 \), Equations (134), (135) and (136) become

\[
\phi_{ij} (X_1, \ldots, X_i; Y_1, \ldots, Y_j) = E \{ \phi (X_1, \ldots, X_a; Y_1, \ldots, Y_a) \mid X_1, \ldots, X_i, Y_1, \ldots, Y_j \} ,
\]  

\[
\sigma_{ij}^2 = \operatorname{Var} \phi_{ij} (X_1, \ldots, X_i; Y_1, \ldots, Y_j) ,
\]  

and

\[
\sigma_{ij}^2 = \operatorname{Cov} \{ \phi (X_1, \ldots, X_a; Y_1, \ldots, Y_a),
\phi (X_1, \ldots, X_i, X_{a_1+1}, \ldots, X_{a_1+(a_1-j)}; Y_1, \ldots, Y_j, Y_{a_2+1}, \ldots, Y_{a_2+(a_2-j)}) \} .
\]
respectively, for $0 \leq i \leq a_1$ and $0 \leq j \leq a_2$.

Although we will not derive it here as we did for the single-sample case, there is an analogous asymptotic normality result for multisample U-statistics, as follows.

**Theorem 36.1** Suppose that for $i = 1, \ldots, s$, $X_{i1}, \ldots, X_{in_i}$, is a random sample from the distribution $F_i$ and that these $s$ samples are independent of each other. Suppose further that there exist constants $\rho_1, \ldots, \rho_s$ in the interval $(0, 1)$ such that $n_i/N \to \rho_i$ for all $i$ and that $\sigma_{a_1 \cdots a_s}^2 < \infty$. Then

$$\sqrt{N}(U_N - \theta) \xrightarrow{d} N(0, \sigma^2),$$

where

$$\sigma^2 = \frac{a_1^2}{\rho_1} \sigma_{10 \cdots 0}^2 + \cdots + \frac{a_s^2}{\rho_s} \sigma_{00 \cdots 0}^2.$$

Although the notation required for the multisample U-statistic theory is nightmarish, life becomes considerably simpler in the case $s = 2$ and $a_1 = a_2 = 1$, in which case we obtain

$$U_N = \frac{1}{n_1n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \phi(X_{1i}; X_{2j}).$$

Equivalently, we may assume that $X_1, \ldots, X_m$ are an iid sample from $F$ and $Y_1, \ldots, Y_n$ are an iid sample from $G$, which gives

$$U_N = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(X_i; Y_j). \quad (137)$$

In the case of the U-statistic of Equation (137), Theorem 36.1 states that

$$\sqrt{N}(U_N - \theta) \xrightarrow{d} N\left(0, \frac{\sigma_{10}^2}{\rho} + \frac{\sigma_{a1}^2}{1 - \rho}\right),$$

where $\rho = \lim m/N$, $\sigma_{10}^2 = \text{Cov} \{\phi(X_1; Y_1), \phi(X_1; Y_2)\}$, and $\sigma_{a1}^2 = \text{Cov} \{\phi(X_1; Y_1), \phi(X_2; Y_1)\}$.

**Example 36.1** For $X_1, \ldots, X_m$ iid from $F$ and $Y_1, \ldots, Y_n$ iid from $G$, consider the Wilcoxon rank-sum statistic $W$, defined to be the sum of the ranks of the $Y_i$ among the combined sample. We showed in Equation (79) that

$$W = \frac{1}{2} n(n + 1) + \sum_{i=1}^{m} \sum_{j=1}^{n} I\{X_i < Y_j\}.$$

Therefore, if we let $\phi(x; y) = I\{x < y\}$, then the corresponding two-sample U-statistic $U_N$ is related to $W$ by $W = \frac{1}{2} n(n + 1) + mnU_N$. Therefore, we may use Theorem 36.1 to obtain the asymptotic normality of $U_N$, and therefore of $W$, just as we did in Topic 26. However, unlike in Topic 26, we make no assumption here that $F$ and $G$ are merely shifted versions of one another. Thus, we may now obtain in principle the asymptotic distribution of the rank-sum statistic for any two distributions $F$ and $G$ that we wish, so long as they have finite second moments.

The other direction in which we will generalize the development of U-statistics is consideration of the joint distribution of two single-sample U-statistics. Suppose that there are two kernel functions, $\phi(x_1, \ldots, x_a)$ and $\varphi(x_1, \ldots, x_b)$, and we define the two corresponding U-statistics

$$U_n^{(1)} = \frac{1}{\binom{n}{a}} \sum_{1 \leq i_1 < \cdots < i_a \leq n} \phi(X_{i_1}, \ldots, X_{i_a})$$

and
and

\[ U_n^{(2)} = \frac{1}{\binom{n}{2}} \sum_{1 \leq j_1 < \cdots < j_n \leq n} \phi(X_{j_1}, \ldots, X_{j_n}) \]

for a single random sample \( X_1, \ldots, X_n \) from \( F \). Define \( \theta_1 = E[U_n^{(1)}] \) and \( \theta_2 = E[U_n^{(2)}] \). Furthermore, define \( \gamma_{ij} \) to be the covariance between \( \phi_i(X_1, \ldots, X_i) \) and \( \phi_j(X_1, \ldots, X_j) \), where \( \phi_i \) and \( \phi_j \) are defined as in Equation (126). It may be proved that

\[ \gamma_{ij} = \text{Cov} \{ \phi(X_1, \ldots, X_a), \varphi(X_1, \ldots, X_j, X_{a+1}, \ldots, X_{a+(b-j)}) \} . \] (138)

Note in particular that \( \gamma_{ij} \) depends only on the value of \( \min\{i, j\} \).

The following theorem, stated without proof, gives the joint asymptotic distribution of \( U_n^{(1)} \) and \( U_n^{(2)} \).

**Theorem 36.2** Suppose \( X_1, \ldots, X_n \) is a random sample from \( F \) and that \( \phi \) and \( \varphi \) are two kernel functions satisfying \( \text{Var} \phi(X_1, \ldots, X_a) < \infty \) and \( \text{Var} \varphi(X_1, \ldots, X_b) < \infty \). Define \( \tau_1^2 = \text{Var} \phi_1(X_1) \) and \( \tau_2^2 = \text{Var} \varphi_1(X_1) \), and let \( \gamma_{ij} \) be defined as in Equation (138). Then

\[ \sqrt{n} \begin{pmatrix} \left( U_n^{(1)} \right) / \left( \theta_1 \right) \\ \left( U_n^{(2)} \right) / \left( \theta_2 \right) \end{pmatrix} \xrightarrow{D} \mathcal{N} \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} a^2 \tau_1^2 & a \bar{b} \gamma_{11} \\ a \bar{b} \gamma_{11} & b^2 \tau_2^2 \end{pmatrix} \right\} . \]

---

### Problems

**Problem 36.1** Suppose \( X_1, \ldots, X_m \) and \( Y_1, \ldots, Y_n \) are independent iid samples from distributions \( \text{Unif}(0, \theta) \) and \( \text{Unif}(\mu, \mu + \theta) \), respectively. Assume \( m/N \to \rho \) as \( m, n \to \infty \) and \( 0 < \mu < \theta \).

(a) Find the asymptotic distribution of the U-statistic of Equation (137), where \( \phi(x, y) = I\{x < y\} \). In so doing, find a function \( g(x) \) such that \( E(U_N) = g(\mu) \).

(b) Find the asymptotic distribution of \( g(\bar{Y} - \bar{X}) \).

(c) Find the range of values of \( \mu \) for which the Wilcoxon estimate of \( g(\mu) \) is asymptotically more efficient than \( g(\bar{Y} - \bar{X}) \). (The ARE in this case is the ratio of asymptotic variances. See the discussion of (4.3.12) and (4.3.13) on p. 240 for details.)

**Problem 36.2** Solve each part of Problem 36.1, but this time under the assumptions that the independent random samples \( X_1, \ldots, X_m \) and \( Y_1, \ldots, Y_n \) satisfy \( P(X_1 \leq t) = P(Y_1 - \theta \leq t) = t^2 \) for \( t \in [0, 1] \) and \( 0 < \theta < 1 \). As in Problem 36.1, assume \( m/N \to \rho \in (0, 1) \).

**Problem 36.3** Suppose \( X_1, \ldots, X_m \) and \( Y_1, \ldots, Y_n \) are independent iid samples from distributions \( \text{Unif}(0, 1) \) and \( \text{N}(\mu, 1) \), respectively. Assume \( m/(m + n) \to 1/2 \) as \( m, n \to \infty \). Let \( U_N \) be the U-statistic of Equation (137), where \( \phi(x, y) = I\{x < y\} \). Suppose that \( \theta(\mu) \) and \( \sigma^2(\mu) \) are such that

\[ \sqrt{N} [U_N - \theta(\mu)] \xrightarrow{D} N[0, \sigma^2(\mu)] . \]

Calculate \( \theta(\mu) \) and \( \sigma^2(\mu) \) for \( \mu \in \{0.2, 0.5, 1, 1.5, 2\} \).

**Hint:** This problem requires a bit of numerical integration. There are a couple of ways you might do this. Mathematica will do it quite easily. There is a function called \texttt{integrate} in Splus and one called \texttt{quad} in MATLAB for integrating a function. If you cannot get any of these to work for you, let me know.
Problem 36.4  Suppose $X_1, X_2, \ldots$ are iid with finite variance. Define

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

and let $G_n$ be Gini's mean difference, the U-statistic defined in Problem 35.2. Note that $S_n^2$ is also a U-statistic, corresponding to the kernel function $\phi(x_1, x_2) = (x_1 - x_2)^2$.

(a) If $X_i$ are distributed as $\text{Unif}(0, \theta)$, give the joint asymptotic distribution of $G_n$ and $S_n$ by first finding the joint asymptotic distribution of the U-statistics $G_n$ and $S_n^2$. Note that the covariance matrix need not be positive definite; in this problem, the covariance matrix is singular.

(b) The singular asymptotic covariance matrix in this problem implies that as $n \to \infty$, the joint distribution of $G_n$ and $S_n$ becomes concentrated on a line. Does this appear to be the case? For 1000 samples of size $n$ from $\text{Uniform}(0, 1)$, plot scatterplots of $G_n$ against $S_n$. Take $n \in \{5, 25, 100\}$. 

107
37. Bootstrapping
Lehmann §6.5

In this topic, we introduce the bootstrap, even though the asymptotics content of what follows is not very high. Essentially, the only large-sample theory we rely on here is the weak law of large numbers. Nonetheless, bootstrapping is included here because of its natural relationship with the concepts of statistical functionals and plug-in estimators seen in recent topics, and also because it is an increasingly popular and often misunderstood method.

Consider a statistical functional \( \lambda_n(F) \) that depends on \( n \). For instance, \( \lambda_n(F) \) may be some property, such as bias or variance, of an estimator \( \hat{\theta}_n \) of \( \theta = \theta(F) \) based on a random sample of size \( n \) from some distribution \( F \).

As an example, let \( \theta(F) = F^{-1}(\frac{1}{2}) \) be the median of \( F \). Take \( \hat{\theta}_n \) to be the \( m \)th order statistic from a random sample of size \( n = 2m - 1 \) from \( F \).

Consider the bias \( \lambda^B_n(F) = E_F \hat{\theta}_n - \theta(F) \) and the variance \( \lambda^V_n(F) = E_F (\hat{\theta}_n - E_F \hat{\theta}_n)^2 \). Theoretical properties of \( \lambda^B_n \) and \( \lambda^V_n \) are very difficult to obtain. Even asymptotics aren’t very helpful, since \( \sqrt{n}(\hat{\theta}_n - \theta) \overset{d}{\approx} N\{0,1/(4f^2(\theta))\} \) tells us only that the bias goes to zero and the limiting variance may be very hard to estimate because it involves the unknown quantity \( f(\theta) \), which is hard to estimate.

Consider the plug-in estimators \( \lambda^B_n(\hat{F}_n) \) and \( \lambda^V_n(\hat{F}_n) \). (Recall that \( \hat{F}_n \) denotes the empirical distribution function, which puts a mass of \( \frac{1}{n} \) on each of the \( n \) sample points.) In our median example,

\[
\lambda^B_n(\hat{F}_n) = E_{\hat{F}_n} \hat{\theta}_n - \hat{\theta}_n
\]

and

\[
\lambda^V_n(\hat{F}_n) = E_{\hat{F}_n} (\hat{\theta}_n)^2 - (E_{\hat{F}_n} \hat{\theta}_n)^2,
\]

where \( \hat{\theta}_n \) is the sample median from a random sample \( X_1^*, \ldots, X_n^* \) from \( \hat{F}_n \).

To see how difficult it is to calculate \( \lambda^B_n(\hat{F}_n) \) and \( \lambda^V_n(\hat{F}_n) \), consider the simplest nontrivial case, \( n = 3 \): Conditional on the order statistics \( (X_1, X_2, X_3) \), there are 27 equally likely possibilities for the value of \( (X_1^*, X_2^*, X_3^*) \), the sample of size 3 from \( \hat{F}_n \), namely

\[
(X(1), X(1), X(1)), (X(1), X(1), X(2)), \ldots, (X(3), X(3), X(3)).
\]

Of these 27 possibilities, exactly 1 + 6 = 7 have the value \( X(1) \) occurring 2 or 3 times. Therefore, we obtain

\[
P(\hat{\theta}_n = X(1)) = \frac{7}{27}, P(\hat{\theta}_n = X(2)) = \frac{13}{27}, \text{ and } P(\hat{\theta}_n = X(3)) = \frac{7}{27}.
\]

This implies that

\[
E_{\hat{F}_n} \hat{\theta}_n^* = \frac{1}{27}(7X(1) + 13X(2) + 7X(3)) \quad \text{and} \quad E_{\hat{F}_n} (\hat{\theta}_n^*)^2 = \frac{1}{27}(7X(1)^2 + 13X(2)^2 + 7X(3)^2).
\]

Therefore, since \( \hat{\theta}_n = X(2) \), we obtain

\[
\lambda^B_n(\hat{F}_n) = \frac{1}{27}(7X(1) - 14X(2) + 7X(3))
\]

and

\[
\lambda^V_n(\hat{F}_n) = \frac{14}{729}(10X(1)^2 + 13X(2)^2 + 10X(3)^2 - 13X(1)X(2) - 13X(2)X(3) - 7X(1)X(3)).
\]
To obtain the sampling distribution of these estimators, of course, we would have to consider the joint
distribution of \((X_1, X_2, X_3)\). Naturally, the calculations become even more difficult as \(n\) increases.

Alternatively, we could use resampling in order to approximate \(\lambda_n^B(\hat{F}_n)\) and \(\lambda_n^V(\hat{F}_n)\). This is the boot-
strapping idea, and it works like this: For some large number \(B\), simulate \(B\) random samples from \(\hat{F}_n\),
namely
\[
X_{11}^*, \ldots, X_{1n}^*, \ldots, X_{B1}^*, \ldots, X_{Bn}^*,
\]
and approximate a quantity like \(E_{\hat{F}_n} \hat{\theta}^*_n\) by the sample mean
\[
\frac{1}{B} \sum_{i=1}^{B} \hat{\theta}^*_i
\]
where \(\hat{\theta}^*_i\) is the sample median of the \(i\)th bootstrap sample \(X_{11}^*, \ldots, X_{in}^*\). Notice that the weak law of large
numbers asserts that
\[
\frac{1}{B} \sum_{i=1}^{B} \hat{\theta}^*_i \xrightarrow{P} E_{\hat{F}_n} \hat{\theta}^*_n.
\]

To recap, then, we wish to estimate some parameter \(\lambda_n(F)\) for an unknown distribution \(F\) based on a random
sample from \(F\). We estimate \(\lambda_n(F)\) by \(\lambda_n(\hat{F}_n)\), but it is not easy to evaluate \(\lambda_n(\hat{F}_n)\) so we approximate
\(\lambda_n(\hat{F}_n)\) by resampling \(B\) times from \(\hat{F}_n\) and obtain a bootstrap estimator \(\lambda_{B,n}^*\). Thus, there are two relevant
issues:

1. How good is the approximation of \(\lambda_n(\hat{F}_n)\) by \(\lambda_{B,n}^*\)? (Note that \(\lambda_n(\hat{F}_n)\) is NOT an unknown parameter;
it is “known” but hard to evaluate.)

2. How precise is the estimation of \(\lambda_n(F)\) by \(\lambda_n(\hat{F}_n)\)?

Question 1 is usually easy to answer using asymptotics; these asymptotics involve letting \(B \to \infty\) and
usually rely on the weak law or the central limit theorem. For example, if we have an expectation functional
\(\lambda_n(F) = E_{\hat{F}_n} h(X_1, \ldots, X_n)\), then
\[
\lambda_{B,n}^* = \frac{1}{B} \sum_{i=1}^{B} h(X_{1i}^*, \ldots, X_{ni}^*) \xrightarrow{P} \lambda_n(\hat{F}_n)
\]
as \(B \to \infty\).

Question 2, on the other hand, is often tricky; asymptotic results involve letting \(n \to \infty\) and are handled
case-by-case. We will not discuss these asymptotics here. On a related note, however, there is an argument
in Lehmann’s book (on pages 432–433) about why a plug-in estimator may be better than an asymptotic
estimator. That is, if it is possible to show \(\lambda_n(F) \to \lambda\) as \(n \to \infty\), then as an estimator of \(\lambda_n(F)\), \(\lambda_n(\hat{F}_n)\)
may be preferable to \(\lambda\).

We conclude this topic by considering the so-called parametric bootstrap. If we assume that the unknown
distribution function \(F\) comes from a family of distribution functions indexed by a parameter \(\mu\), then \(\lambda_n(F)\)
is really \(\lambda_n(\hat{F}_\mu)\). Then, instead of the plug-in estimator \(\lambda_n(\hat{F}_\mu)\), we might consider the estimator \(\lambda_n(\hat{F}_\hat{\mu})\),
where \(\hat{\mu}\) is an estimator of \(\mu\).

Everything proceeds as in the nonparametric version of bootstrapping. Since it may not be easy to
evaluate \(\lambda_n(\hat{F}_\hat{\mu})\) explicitly, we first find \(\hat{\mu}\) and then take \(B\) random samples of size \(n\), \(X_{11}^*, \ldots, X_{in}^*\) through
\(X_{B1}^*, \ldots, X_{Bn}^*\), from \(\hat{F}_\hat{\mu}\). These samples are used to approximate \(\lambda_n(\hat{F}_\hat{\mu})\).
Example 37.1  Suppose $X_1, \ldots, X_n$ is a random sample from Poisson($\mu$). Take $\hat{\mu} = \bar{X}$. Suppose $\lambda_n(F_{\hat{\mu}}) = \text{Var } F_{\hat{\mu}}$. In this case, we happen to know that $\lambda_n(F_{\hat{\mu}}) = \mu/n$, but let’s ignore this knowledge and apply a parametric bootstrap. For some large $B$, say 500, generate $B$ samples from Poisson($\hat{\mu}$) and use the sample variance of $\hat{\mu}^*$ as an approximation to $\lambda_n(F_{\hat{\mu}})$. In R, with $\mu = 1$ and $n = 20$ we obtain

```r
> x <- rpois(20,1) # Generate the sample from F
> muhat <- mean(x)
> muhat
[1] 0.85
> muhatstar <- rep(0,500) # Allocate the vector for muhatstar
> for(i in 1:500) muhatstar[i] <- mean(rpois(20,muhat))
> var(muhatstar)
[1] 0.04139177
```

Note that the estimate 0.041 is close to the known true value 0.05. Obviously, this example is simplistic because we already know that $\lambda_n(F) = \mu/n$, which makes $\hat{\mu}/n$ a more natural estimator. However, it is not always so simple to obtain a closed-form expression for $\lambda_n(F)$.

Incidentally, we could also use a nonparametric bootstrap approach in this example:

```r
> for (i in 1:500) muhatstar2[i] <- mean(sample(x,replace=T))
> var(muhatstar2)
[1] 0.0418454
```

Of course, 0.042 is an approximation to $\lambda_n(F_n)$ rather than $\lambda_n(F_{\hat{\mu}})$. Furthermore, we can obtain a result arbitrarily close to $\lambda_n(F_n)$ by increasing the value of $B$:

```r
> muhatstar2_rep(0,100000)
> for (i in 1:100000) muhatstar2[i] <- mean(sample(x,replace=T))
> var(muhatstar2)
[1] 0.04136046
```

In fact, it is in principle possible to obtain an approximate variance for our estimates of $\lambda_n(F_n)$ and $\lambda_n(F_{\hat{\mu}})$, and, using the central limit theorem, construct approximate confidence intervals for these quantities. This would allow us to specify the quantities to any desired level of accuracy.

---

Problems

Problem 37.1  (a) Devise a nonparametric bootstrap scheme for setting confidence intervals for $\beta$ in the linear regression model $Y_i = \alpha + \beta x_i + \epsilon_i$. There is more than one possible answer.

(b) Using $B = 1000$, implement your scheme on the following dataset to obtain a 95% confidence interval. Compare your answer with the standard 95% confidence interval.

<table>
<thead>
<tr>
<th>$Y$</th>
<th>21</th>
<th>16</th>
<th>20</th>
<th>34</th>
<th>33</th>
<th>43</th>
<th>47</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>460</td>
<td>498</td>
<td>512</td>
<td>559</td>
<td>614</td>
<td>675</td>
<td>719</td>
</tr>
</tbody>
</table>

(In the dataset, $Y$ is the number of manatee deaths due to collisions with powerboats in Florida and $x$ is the number of powerboat registrations in thousands for even years from 1978-1990.)

Problem 37.2  Consider the following dataset that lists the latitude and mean August temperature for 7 US cities. The residuals are listed for use in part (b).
Minitab gives the following output for a simple linear regression:

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>113.99</td>
<td>13.01</td>
<td>8.76</td>
<td>0.000</td>
</tr>
<tr>
<td>latitude</td>
<td>-0.9730</td>
<td>0.3443</td>
<td>-2.83</td>
<td>0.037</td>
</tr>
</tbody>
</table>

\[
S = 5.546 \quad \text{R-Sq = 61.5\%} \quad \text{R-Sq(adj) = 53.8\%}
\]

Note that this gives an asymptotic estimate of the variance of the slope parameter as \(.3443^2 = 0.1185\).

In each case below, use the described method to simulate \(B = 500\) bootstrap samples \((x_{bi}^*, y_{bi}^*)\) for \(b = 1, \ldots, B\). For each \(b\), refit the model to obtain \(\hat{\beta}_b^*\). Report the sample variance of \(\hat{\beta}_0^*, \ldots, \hat{\beta}_B^*\) and compare with the asymptotic estimate of \(0.1185\).

(a) **Parametric bootstrap.** Take \(x_{bi}^* = x_i\) for all \(b\) and \(i\). Let \(y_{bi}^* = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i\), where \(\epsilon_i \sim N(0, \hat{\sigma}^2)\). Obtain \(\hat{\beta}_0, \hat{\beta}_1\), and \(\hat{\sigma}^2\) from the above output.

(b) **Nonparametric bootstrap I.** Take \(x_{bi}^* = x_i\) for all \(b\) and \(i\). Let \(y_{bi}^* = \hat{\beta}_0 + \hat{\beta}_1 x_i + r_{bi}^*\), where \(r_{b1}^*, \ldots, r_{b7}^*\) is an iid sample from the empirical distribution of the residuals from the original model (you may want to refit the original model to find these residuals).

(c) **Nonparametric bootstrap II.** Let \((x_{b1}^*, y_{b1}^*), \ldots, (x_{b7}^*, y_{b7}^*)\) be an iid sample from the empirical distribution of \((x_1, y_1), \ldots, (x_7, y_7)\).

**Note:** In Splus, you can obtain the slope coefficient of the linear regression of the vector \(y\) on the vector \(x\) using \(\text{lm}(y \sim x)$coef[2]$\).

**Problem 37.3** The same resampling idea that is exploited in the bootstrap can be used to approximate the value of difficult integrals by a technique sometimes called Monte Carlo integration. Suppose we wish to compute

\[
\theta = 2 \int_0^1 e^{-x^2} \cos^3(x) \, dx.
\]

Using the \texttt{NIntegrate} function in Mathematica, we find that \(\theta = 1.07051569222\).

(a) Define \(g(t) = 2e^{-t^2} \cos^3(t)\). Let \(U_1, \ldots, U_n\) be an iid uniform(0,1) sample. Let

\[
\hat{\theta}_1 = \frac{1}{n} \sum_{i=1}^{n} g(U_i).
\]

Prove that \(\hat{\theta}_1 \overset{p}{\rightarrow} \theta\).

(b) Define \(h(t) = 2 - 2t\). Prove that if we take \(V_i = 1 - \sqrt{U_i}\) for each \(i\), then \(V_i\) is a random variable with density \(h(t)\). Prove that with

\[
\hat{\theta}_2 = \frac{1}{n} \sum_{i=1}^{n} \frac{g(V_i)}{h(V_i)},
\]

we have \(\hat{\theta}_2 \overset{p}{\rightarrow} \theta\).
(c) For $n = 1000$, simulate $\hat{\theta}_1$ and $\hat{\theta}_2$. Give estimates of the variance for each estimator by reporting $\hat{\sigma}^2/n$ for each, where $\hat{\sigma}^2$ is the sample variance of the $g(U_i)$ or the $g(V_i)/h(V_i)$ as the case may be.

(d) Plot, on the same set of axes, $g(t)$, $h(t)$, and the standard uniform density for $t \in [0,1]$. From this plot, explain why the variance of $\hat{\theta}_2$ is smaller than the variance of $\hat{\theta}_1$. [Incidentally, the technique of drawing random variables from a density $h$ whose shape is close to the function $g$ of interest is a variance-reduction technique known as importance sampling.]

Note: This was sort of a silly example, since it was really easy to get an exact value for $\theta$ using numerical methods. However, with certain high-dimensional integrals, the “curse of dimensionality” makes exact numerical methods extremely time-consuming computationally; thus, Monte Carlo integration does have a practical use in such cases.