1978

**Estimation of $\mu_y$ Using the General Regression Model (in sampling)**

Michael R. Manieri

*Utah State University*

---

Follow this and additional works at: [https://digitalcommons.usu.edu/gradreports](https://digitalcommons.usu.edu/gradreports)

Part of the [Applied Statistics Commons](https://digitalcommons.usu.edu/gradreports)

---

**Recommended Citation**


[https://digitalcommons.usu.edu/gradreports/1185](https://digitalcommons.usu.edu/gradreports/1185)

---

This Report is brought to you for free and open access by the Graduate Studies at DigitalCommons@USU. It has been accepted for inclusion in All Graduate Plan B and other Reports by an authorized administrator of DigitalCommons@USU. For more information, please contact digitalcommons@usu.edu.
ESTIMATION OF $\mu_Y$ USING THE GENERAL REGRESSION MODEL
(in sampling)

by

Michael R. Manieri

A report submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE in

Applied Statistics

Plan B

Approved:

UTAH STATE UNIVERSITY
Logan, Utah

1978
ACKNOWLEDGEMENTS

I would like to especially thank my major professor, Dr. David L. Turner, for the many hours of precious help he extended to me during my studies. As a friend and teacher he made this time at Utah State both enjoyable and educational.

Thanks also to my committee members Dr. Ronald V. Canfield and Dr. Gregory W. Jones.

I would also like to thank the Kolesar's and Relli's, whose generosity enabled me to stay at their homes during the writing of this report.

I dedicate this work to my parents, who helped me both financially, and with their love and encouragement throughout my studies. Also to my fiancée Lidia, whose love and prayers guided me through this past year.

I also thank God for this opportunity to serve Him. I pray that I can always share the love that I have received during this past year.

Michael R. Manieri
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>ii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>v</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>vi</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>RATIO ESTIMATORS</td>
<td>3</td>
</tr>
<tr>
<td>SIMPLE LINEAR REGRESSION ESTIMATORS</td>
<td>12</td>
</tr>
<tr>
<td>THE GENERAL REGRESSION MODEL</td>
<td>17</td>
</tr>
<tr>
<td>CONCLUSION</td>
<td>24</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>29</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table

2.1. 10 nail orders and their corresponding weights (in lbs.).

2.2. Age vs. eye lens weight for road killed deer.

3.1. Blood hemoglobin for dogs as the percentage of normal cells and red blood cells in millions per cubic millimeter.

5.1. Results obtained using the models and variance techniques discussed for the deer example.
LIST OF FIGURES

Figure

2.1. Plot of pounds vs. number of nails for data in Table 2.1.

2.2. Plot of eye lens weight vs. age for road killed deer.

3.1. Plot of blood hemoglobin as the percentage of normal cells and red blood cells in millions per cubic millimeter, for dogs.
ABSTRACT

Estimation of $\mu_Y$ using the General Regression Model (in sampling)

by

Michael R. Manieri, Master of Science
Utah State University, 1978

Major Professor: Dr. David L. Turner
Department: Applied Statistics

The methods of ratio and regression estimators discussed by Cochran (1977) are given as background materials and extended to the estimation of $\mu_Y$, the population mean of the $Y$'s, using a general regression model.

The propagation of error technique given by Deming (1948) is used as an approximation to find the variance of the estimator $\hat{\mu}_Y$.

Examples are given for each of the various models. Variances of $\hat{\mu}_Y$ are calculated and compared.

(29 pages)
CHAPTER I
INTRODUCTION

One of the most common and elementary concepts in statistics is that of regression. In sampling problems, we can often improve the precision of an estimator of $\mu_Y$ by using one or more independent variables, say $X_1, X_2, \ldots, X_k$.

Cochran (1977), discusses only ratio and simple linear regression estimators which are restricted to a single $X$ variable. For these cases, Cochran (1977) develops procedures for estimating $\mu_Y$ and gives approximate standard errors of these estimates.

As is often the case, we wish to generalize our model to where we can fit all types of data and more general models. We can use regression techniques to get our estimates of $\mu_Y$ but the variance of $\hat{\mu}_Y$ can often be very difficult to compute.

In the following pages, approximations for $\text{Var}(\hat{\mu}_Y)$, which can be easily computed, are derived. With these approximations, we can use the best model for the data and are not restricted to the use of ratio and regression estimates.

In this report we wish to generalize Cochran's (1977) ratio and regression techniques to allow us to compute estimates of $\mu_Y$ and approximations for the variance of these estimators.

In Chapter II we review Cochran's (1977) ratio estimation procedure and then develop a regression model without an
intercept. Chapter III covers the simple linear regression model while Chapter IV is devoted to the general regression model.

Each technique will be used on at least the sample set of data described in Table 2.2.
The basic purpose of using an additional variable $X$ is that, because of its correlation with $Y$, we are able to increase the precision of our estimating procedures. This increase in precision is measured by a decrease in the variance of the estimators with a corresponding decrease in the width of the confidence intervals.

The simplest case is that of the ratio estimate. Here we assume that the $Y$ variable is directly proportional to a single $X$ variable. We assume that the population ratio, $\rho = \mu_Y/\mu_X$, is unknown, but that $\mu_X$ is known. From a sample, we then compute an estimate of $\rho$, namely $R = \sum Y_i/\sum X_i$, giving us the following estimate of $\mu_Y$:

$$\hat{\mu}_Y = \left[\frac{\sum Y_i}{\sum X_i}\right] \mu_X \quad [2.1]$$

An example at this point may help to demonstrate this concept. Suppose we run a wholesale hardware store which sells nails as one of its main products. Our customers could request anywhere from 1,000 to 25,000 or more nails. Rather than count out the exact number of nails we could use ratio estimators to estimate the number of nails by measuring the weight of the nails. As our sample we took 10 random orders...
from the 114 made during the previous month, counted out the nails and then obtained the weight of the nails for these orders. \( \mu_X \) is assumed to be 850 pounds which is the population mean of the 114 individual orders. The data for this problem is given in Table 2.1.

For this example we see that our estimate of \( p \) is \( r = 14.2115 \) which gives us an estimate of 12079.775 for \( \mu_Y \).

We can observe that our plot is fairly linear and goes through the origin. Since zero weight implies no nails, ratio estimation is an appropriate method to use. The idea here is that it is easier and quicker to measure the weight of the nails than it is to count out each order.

After getting this estimate we need to be able to know how good our estimate is. The variance of the estimator helps us do this. It allows us to set confidence limits around the parameter. Cochran (1977) shows the estimated variance of \( \hat{\beta}_Y \) to be approximately

\[
\text{Var}(\hat{\beta}_Y) = \frac{N-n}{Nn} \left[ \frac{\sum (Y_i - RX_i)^2}{n-1} \right]
\]

which can be more easily computed by the following formula:

\[
\text{Var}(\hat{\beta}_Y) = \frac{N-n}{Nn} \left[ \frac{\sum Y_i^2 + R^2 \sum X_i^2 - 2RX_i \sum Y_i}{n-1} \right]
\]

The approximation is necessary since the exact variance of a ratio of random variables is not easy to work with. For the
Table 2.1. 10 nail orders and their corresponding weights (in lbs.).

<table>
<thead>
<tr>
<th>X = weight of nails</th>
<th>Y = number of nails</th>
</tr>
</thead>
<tbody>
<tr>
<td>84.4</td>
<td>1200</td>
</tr>
<tr>
<td>126.7</td>
<td>1800</td>
</tr>
<tr>
<td>190.2</td>
<td>2700</td>
</tr>
<tr>
<td>422.9</td>
<td>6000</td>
</tr>
<tr>
<td>845.6</td>
<td>12000</td>
</tr>
<tr>
<td>1055.8</td>
<td>15000</td>
</tr>
<tr>
<td>1264.9</td>
<td>18000</td>
</tr>
<tr>
<td>1402.5</td>
<td>20000</td>
</tr>
<tr>
<td>1585.5</td>
<td>22500</td>
</tr>
<tr>
<td>1760.9</td>
<td>25000</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\sum X_i &= 8739.4 & \sum Y_i &= 124200 \\
\sum X_i^2 &= 11249507.02 & \sum Y_i^2 &= 2272220000 \\
\sum X_iY_i &= 159878930
\end{align*}
\]
nail example we get the following results:

\[
\text{Var}(\hat{\beta}_Y) = \frac{104}{114\cdot 10} \left[ \frac{7344}{9} \right] = 74.442
\]  

An approximate (1-\(\alpha\)) confidence interval for \(\mu_Y\) would then be:

\[
\hat{\beta}_Y \pm Z_{\alpha/2} \sqrt{\text{Var}(\hat{\beta}_Y)}
\]  

For the nail example, this 95% confidence interval is (12062.87, 12096.69).

The model implied by the ratio estimator is \(Y = \rho X\) which may be interpreted as a regression model which passes through the origin. This leads us to try regression theory to generate another estimator. For our model, we assume \(Y = \beta_1 X + \varepsilon\) where the \(\varepsilon\) is the stochastic or random part of the model. The model still passes through the origin, but our estimator of \(\beta_1\) is different from the estimate we got using ratio techniques.

It is easily shown that the least squares estimate of \(\beta_1\) is

\[
\hat{\beta}_1 = \frac{\Sigma X_i Y_i}{\Sigma X_i^2}.
\]  

This leads to an estimate of \(\mu_Y\) as

\[
\hat{\mu}_Y = \hat{\beta}_1 \mu_X
\]  

We can compare the results of this estimator with the ratio estimator by comparing variances for the \(\hat{\beta}_Y\)'s. If we assume that the \(X\) variable is fixed, Graybill(1977) gives results
which show the estimated variance of $\hat{\mu}_Y$ to be

$$\text{Var}(\hat{\mu}_Y) = S^2 \left[ \frac{1}{n} + \frac{(\mu_X-\bar{X})^2}{\text{EX}_i^2} + \frac{2(\hat{\beta}_1 \bar{X} - \bar{Y})}{\text{EX}_i} \right]$$

[2.8]

where $S^2$ is $[\sum Y_i^2 - ((\sum X_i Y_i)/\sum X_i)^2]/(n-1)$.

Figure 2.1 is a plot of the nail data given in Table 2.1. From this plot it is evident that the relationship is linear and does pass through the origin, which implies that the no intercept regression model is also appropriate. This gives us an estimated slope, $\hat{\beta}_1 = 14.212$. Using this value of $\hat{\beta}_1$ we can apply formula 2.7 to get an estimate of $\mu_Y$ as $\hat{\mu}_Y = 12080.2$. Formula 2.8 can then be used to get an estimate of the variance of $\mu_Y$ as 77.15 which gives a 95% confidence interval of (12062.98 , 12097.42). Note that this confidence interval is wider than the one given by the ratio estimator.

As a second example, consider the data presented in Table 2.2 where the eye lens weights of road killed deer were measured as the Y variable. The X variable was taken as the age in months of the deer. It seems clear from Figure 2.3 that neither the ratio nor the no intercept regression model is exactly right for this set of data since the relationship does not appear to be linear. It does appear to pass through the origin however.

Use of the ratio estimator on this set of data gives an estimate of $\mu_Y = 1.28$ with a confidence interval of (.888 , 1.67). If we use the no intercept regression estimator, we get an
Figure 2.1. Plot of pounds vs. number of nails for data in Table 2.1.
Table 2.2. Age vs. eye lens weight for road killed deer.

<table>
<thead>
<tr>
<th>X = age</th>
<th>Y = eye lens weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>83</td>
<td>1.26</td>
</tr>
<tr>
<td>12</td>
<td>.88</td>
</tr>
<tr>
<td>3</td>
<td>.50</td>
</tr>
<tr>
<td>41</td>
<td>1.25</td>
</tr>
<tr>
<td>32</td>
<td>1.14</td>
</tr>
<tr>
<td>9</td>
<td>.74</td>
</tr>
<tr>
<td>24</td>
<td>1.07</td>
</tr>
<tr>
<td>73</td>
<td>1.32</td>
</tr>
<tr>
<td>15</td>
<td>.83</td>
</tr>
<tr>
<td>29</td>
<td>1.03</td>
</tr>
</tbody>
</table>

\[ \Sigma X_1 = 321 \quad \Sigma Y_1 = 10.02 \]
\[ \Sigma X_1^2 = 16799 \quad \Sigma Y_1^2 = 10.6588 \]
\[ \Sigma X_1 Y_1 = 375.39 \]
Figure 2.2. Plot of eye lens weight vs. age for road killed deer.
estimate of \( \mu_y = 0.916 \) and a corresponding 95% confidence interval of \((0.522, 1.31)\).

The question naturally arises as to which of these two estimators is best. Note that the ratio estimator is derived from a line passing through the origin and the point \((\bar{X}, \bar{Y})\). The no intercept regression model computes its slope differently using \( \hat{\beta}_1 = \frac{\sum X_i Y_i}{\sum X_i^2} \).

Mendenhall (1971) and Cochran (1977) both suggest the use of ratio estimators when the relationship between \( Y \) and \( X \) is linear and goes through the origin and when the variance of \( Y \) about this line is proportional to \( X \).

In the next Chapter an intercept will be added. This simple regression estimator can be shown to do at least as good or better than the no intercept model.
CHAPTER III

SIMPLE LINEAR REGRESSION ESTIMATORS

Like ratio estimators, we are using a single auxiliary variable to help us improve precision in estimating \( \mu_Y \). The linear regression of \( Y \) on \( X \), which now does not have to pass through the origin, is the basis of this estimator. Our model is \( Y = \beta_0 + \beta_1 X + \epsilon \) which leads to the linear regression estimate of \( \mu_Y \) as

\[
\hat{\beta}_Y = \hat{\beta}_0 + \hat{\beta}_1 \mu_X
\]  \[3.1\]

As our estimates of \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \), we take the usual least squares estimators,

\[
\hat{\beta}_1 = \frac{\Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)/n}{\Sigma X_i^2 - (\Sigma X_i)^2/n}
\]  \[3.2\]

\[
\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}
\]  \[3.3\]

The following example should help to give a clear presentation of regression estimates. The example data was taken from Dixon and Massey (1969).

An experiment was performed to measure blood hemoglobin in dogs as to the percentage of normal cells and red blood cells in millions per cubic millimeter. The data is given in Table 3.1. \( \mu_X \) is given as 94. Figure 3.1 is a plot of the data in Table 3.1.
Table 3.1. Blood hemoglobin as the percentage of normal cells and red blood cells in millions per cubic millimeter, for dogs.

<table>
<thead>
<tr>
<th>$X = \text{blood hemoglobin}$</th>
<th>$Y = \text{red blood cells}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>93</td>
<td>7.3</td>
</tr>
<tr>
<td>96</td>
<td>6.5</td>
</tr>
<tr>
<td>108</td>
<td>7.7</td>
</tr>
<tr>
<td>86</td>
<td>5.4</td>
</tr>
<tr>
<td>92</td>
<td>6.7</td>
</tr>
<tr>
<td>80</td>
<td>5.1</td>
</tr>
<tr>
<td>84</td>
<td>7.2</td>
</tr>
<tr>
<td>117</td>
<td>8.5</td>
</tr>
<tr>
<td>95</td>
<td>7.8</td>
</tr>
<tr>
<td>94</td>
<td>6.6</td>
</tr>
</tbody>
</table>

$\Sigma X_1 = 945$  $\Sigma Y_1 = 68.8$

$\Sigma X_1^2 = 90395$  $\Sigma Y_1^2 = 483.38$

$\Sigma X_1 Y_1 = 6584$
Figure 3.1. Plot of blood hemoglobin as the percentage of normal cells and red blood cells in millions per cubic millimeter, for dogs.
Note that the relationship is roughly linear.

Using formula 3.2 and 3.3 we calculate $\hat{\beta}_1 = .0754$ and $\hat{\beta}_0 = -.2453$. Our prediction equation is given by $Y = -.2453 + .0754X$.

The estimate of $\mu_Y$, using these estimates is then 6.7669.

If we use the regression techniques discussed in Graybill(1977) or Draper and Smith(1966), we can show the estimated variance of $\hat{\beta}_Y$ to be

$$\tilde{\text{Var}}(\hat{\beta}_Y) = S^2 \left[ \frac{1}{n} + \frac{(\mu_X - \bar{X})^2}{\left( \sum X_i^2 - (\sum X_i)^2/n \right)} \right] \quad [3.4]$$

where $S^2$ is given as

$$S^2 = \left[ \sum Y_i^2 - (\sum Y_i)^2/n - \frac{\sum X_i Y_i - (\sum X_i)(\sum Y_i)^2/n}{\sum X_i^2 - (\sum X_i)^2/n} \right] \frac{1}{n-2} \quad [3.5]$$

Cochran(1977) gives the estimated variance of $\hat{\beta}_Y$ as

$$\tilde{\text{Var}}(\hat{\beta}_Y) = \left[ \frac{1}{n(n-2)} \right] \left[ \frac{n\sum Y_i^2 - (\sum Y_i)^2 - (n\sum X_i Y_i - \sum X_i \sum Y_i)^2}{n(\sum X_i^2) - (\sum X_i)^2} \right] \quad [3.6]$$

which is a different way of writing $S^2$ given by equation 3.5.

Equation 3.4 seems more intuitively appealing since it takes the sample $\bar{X}$ into account. If $(\mu_X - \bar{X})$ is as large or larger than the sum of squares for $X$, then the estimated variance given by formula 3.4 would be larger than that given by formula 3.6. Usually this is not the case unless $\mu_X$ is very far from $\bar{X}$.

For the hemoglobin example Cochran's(1977) $\tilde{\text{Var}}(\hat{\beta}_Y) = .4776$ while formula 3.4 gives $\tilde{\text{Var}}(\hat{\beta}_Y) = .04787$. Cochran's(1977) 95% confidence interval would then be (5.41, 8.12) and using
Graybill (1977) we get (6.34, 7.20). We obtain a narrower confidence interval, as suggested, using the variance computed by formula 3.4.

Since the intercept is close to zero, the ratio and no intercept regression models were also tried. The estimate of \( \mu_Y \) for the ratio estimator is 6.77, with a 95% confidence interval of (6.23, 7.31).

Using the no intercept model, we get an estimate of 6.77 for \( \mu_Y \) and a confidence interval of (6.37, 7.17). Note that for this data a shorter confidence interval was obtained using the no intercept model.

As a second example for simple regression estimators, the deer data presented in Table 2.2 will be used. Using formula 3.2 and 3.3, \( b_1 \) is calculated to be 0.008775 and \( b_0 \) is 0.736. Using these two values in equation 3.1 we get \( \mu_Y = 1.0757 \).

Using equation 3.4 we get \( \text{Var}(b_Y) = 0.00220 \) with a 95% confidence interval of (0.987, 1.168). Using Cochran's (1977) formula 3.6 we get the \( \text{Var}(\hat{\mu}_Y) = 0.0217 \) with a confidence interval of (0.787, 1.364).

We can see by these results that our model with an intercept increases the precision of our estimator by the narrower confidence interval obtained.
CHAPTER IV

THE GENERAL REGRESSION MODEL

Just as the no intercept or the ratio models were not ideal for all situations, the simple regression model may not have sufficient flexibility for all situations. These models may be generalized to a general linear model or to an even more general regression model.

For the general regression model, we consider

\[ Y = f(\underline{X}; \underline{\beta}) + \epsilon \]  \hspace{1cm} [4.1]

where \( \underline{\beta} \) is a kxl vector of parameters and \( \underline{X} \) is a pxl vector of independent variables whose population means, \( \mu_1, \mu_2, \ldots, \mu_p \), are assumed known. We assume the user has obtained estimate of \( \underline{\beta} \), say \( \hat{\underline{\beta}} \) by some procedure. If the model is linear, procedures given by Graybill(1977) can be followed. If the model is non-linear, procedures sketched by Draper and Smith(1966) or others may be followed.

We let \( f(\underline{X}; \underline{\beta}) \) be our sample estimate of the model given in equation 4.1. As our estimate of \( \mu_Y \) we then take

\[ \hat{\beta}_Y = f(\underline{\mu}; \hat{\underline{\beta}}) \]  \hspace{1cm} [4.2]

After getting our estimate for \( \mu_Y \), the estimated \( \text{Var}(\mu_Y) \) has to be obtained. In the general regression case the exact computational variance often proves to be too difficult for
complex models or when the variables are not normally distributed. We will approximate the variance by using a truncated Taylor series expansion as suggested by Deming (1948).

The variance of the function $f(X)$ can be shown to be

$$
\sigma_f^2 = E[f(X) - E(f(X))]^2
$$

[4.3]

Deming approximates this by the equation

$$
\sigma_f^2 \approx E[f(X) - f(E(X))]^2
$$

[4.4]

where $E[f(X)]$ has been replaced by $f(E(X))$. In practice, $\Sigma f(X_i)/n$ should be compared to $f(\Sigma X_i/n)$ to get some idea of how well this approximation works. If this value is "too" large we will suspect corresponding error in the variance approximation. As an analytical bound, Jensen's inequality as stated by Loeve says that if $f(X)$ is a convex function then $f[E(X)] \leq E[f(X)]$ which makes the approximation (4.4) larger than expected. If the function is concave the inequality is reversed.

We can then apply a Taylor series expansion to $[f(X) - f(E(X))] = \Delta f(X)$. By this expansion we get

$$
\Delta f(X) = \frac{\partial f}{\partial X} \Delta X + \frac{\partial^2 f}{\partial X^2} \frac{\Delta X^2}{2!} + \cdots + \frac{\partial^{n-1} f}{\partial X^{n-1}} \frac{\Delta X^{n-1}}{(n-1)!} + R_n
$$

[4.5]

where the remainder is given by

$$
R_n = \frac{\partial^n f}{\partial X^n} \frac{\Delta X^n}{n!}
$$

[4.6]
\( R_n \) is evaluated at \( X_o \) where \( X_o \) is between \( f(E(X)) \) and \( E[f(X)] \).

Truncation after the first term gives us a good approximation in most cases but more terms can be added if needed.

\[
\sigma_f^2 = E[\Delta f(X)]^2
= E \left[ \frac{\partial f}{\partial X} \Delta X \right]^2
= \frac{\partial f}{\partial X_1} \sigma_{\bar{X}_1}^2 + \frac{\partial f}{\partial X_2} \sigma_{\bar{X}_2}^2 + \ldots + \frac{\partial f}{\partial X_k} \sigma_{\bar{X}_k}^2
\]

[4.7]

When \( X_1, X_2, \ldots, X_k \) are independent random variables then the cross product terms are zero since independent random variables have \( \rho_{X_iX_j} = 0 \). This technique is referred to by Deming (1948) as the propagation of error.

This gives us a workable approximation for the \( \text{Var}(\beta_Y) \) to be used with the general regression model. As before, we can get a \((1-\alpha)\) confidence interval by using equation 2.5.

The exact variance for \( \beta_Y \) for the simple regression model was given by equation 3.4. If we apply equation 4.7 to get an approximation for the variance of \( \beta_Y \), we get close to the expression given by Cochran (1977) for his regression estimator.

\[
\text{Var}(\beta_Y) = \beta_1^2 \sigma_{\bar{X}}^2 + \sigma_{\bar{Y}}^2 + 2(\beta_1)(\sigma_X \sigma_Y) \rho_{XY} \quad [4.8]
\]

Since the population quantities \( \sigma_{\bar{X}}, \sigma_{\bar{Y}}, \) and \( \rho_{XY} \) are generally not known, we use their corresponding sample estimates which gives us the sample estimate of the variance of \( \beta_Y \) as
\[
\text{Var}(\hat{\beta}_Y) = \left[ \frac{\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i / n}{\Sigma X_i^2 - (\Sigma X_i)^2 / n} \right] \frac{1}{n-1} + \left[ \frac{\Sigma Y_i^2 - (\Sigma Y_i)^2}{n} \right] \frac{1}{n-1} \\
- 2 \left[ \frac{(\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i / n)^2}{\Sigma X_i^2 - (\Sigma X_i)^2 / n} \right] \\
= \left[ \frac{n \Sigma Y_i^2 - (\Sigma Y_i)^2 - (n \Sigma X_i Y_i - \Sigma X_i \Sigma Y_i)^2}{n \Sigma X_i^2 - (\Sigma X_i)^2} \right] \frac{1}{n-1} \]  

(Cochran (1977) suggests using \( (n - 2) \) instead of \( (n - 1) \) since it is used in standard regression theory and is known to give an unbiased estimate of \( \text{Var}(\hat{\beta}_Y) \). With this we see that 4.9 corresponds exactly with the \( \text{Var}(\hat{\beta}_Y) \) given in equation 3.6.)

For the regression model without an intercept the \( \text{Var}(\hat{\beta}_Y) \) was given by equation 2.8. If \( X \) is a random variable we could use the propagation of error technique to get an estimate of \( \text{Var}(\hat{\beta}_Y) \) for the model \( Y = \beta_1 X + \epsilon \) where \( \mu_Y = \beta_1 \mu_X \).

\[
\text{Var}(\hat{\beta}_Y) = \left[ \frac{\partial f}{\partial \beta} \right]^2 \sigma^2_x \\
= (-\beta_1)^2 \sigma^2 + \sigma^2_y + 2 (-\beta_1) \sigma_x \sigma_y \rho_{XY} \]  

\[
= \frac{1}{n(n-1)} \left[ \frac{\Sigma X_i Y_i}{\Sigma X_i^2} \left[ \frac{\Sigma X_i Y_i}{\Sigma X_i^2} \left( n \Sigma X_i^2 / (\Sigma X_i)^2 - (\Sigma X_i)^2 / n \right) \right. \\
+ \left. 2 \Sigma X_i Y_i \right] \right] + n \Sigma Y_i^2 - (\Sigma Y_i)^2 \]  

Again if we replace the unknown \( \sigma_x \), \( \sigma_y \) and \( \rho_{XY} \) with their corresponding sample estimates, we come up with the estimate...
of \( \text{Var}(\beta_Y) \) given by equation 4.11.

Computing this \( \text{Var}(\beta_Y) \) for the deer data in Table 2.2 we get \( \text{Var}(\beta_Y) = 0.3649 \). This is a more conservative value for the \( \text{Var}(\beta_Y) \) as compared with the value computed from equation 2.8 which was 0.0404 for the deer example. Here though we have no assumptions on our variables or their distribution.

To illustrate the general regression technique, two models were suggested as of being of interest for the deer data:

1) \( Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon \) and \[4.12\]
2) \( \ln(Y) = \alpha_0 + \alpha_1/(X+7) + \varepsilon \). \[4.13\]

Using the deer data given in Table 2.2, the estimates for \( \beta_0, \beta_1, \beta_2 \) for the polynomial model were as follows:

\[
\hat{\beta}_0 = 0.486 \\
\hat{\beta}_1 = 0.0277 \\
\hat{\beta}_2 = 0.000223
\] \[4.14\]

The estimate of \( \mu_Y \) computed from equation 4.2 is 1.2468. The estimated variance of \( \hat{\beta}_Y \) computed from equation 4.7 is

\[
\text{Var}(\hat{\beta}_Y) = (\hat{\beta}_1 + 2\hat{\beta}_2 \mu_X)^2 \delta_X^2 + \delta_Y^2 + 2(\hat{\beta}_1 + 2\hat{\beta}_2 \mu_X) \delta_X \delta_Y \delta_{X\hat{\beta}_{XY}} \] \[4.15\]

\[
= 0.18891
\]

This estimated variance gives us an approximate 95% confidence interval of \((0.3949, 2.0987)\).
If we assume that the X variables are fixed non-random variables or that the regression of Y on X is linear in the $\beta$ coefficients, we can use results given by Graybill (1977) for calculation of $\text{Var}(\hat{\beta}_Y)$. For full details, see Graybill (1977), but for this example, $\text{Var}(\hat{\beta}_Y) = 0.03686$, giving a confidence interval of $(0.871, 1.623)$.

This estimated variance is considerably smaller than the estimated variance using the propagation of error technique, but we must keep in mind the fact that the propagation of error technique is only an approximation.

The model $\ln(Y) = a_0 + a_1/(x+7) + \varepsilon$ seems to give a good fit to the data. Note that $X$ is the deer's age in months and $7$ is the number of months in a deer's gestation period. We again use the same calculation to estimate $\mu_Y$ and to find $\text{Var}(\hat{\mu}_Y)$. Using regression techniques to compute $a_0, a_1$, we get

$$\hat{a}_0 = 0.391 \quad \hat{a}_1 = -10.9 \quad [4.16]$$

The estimate of $\mu_Y$ computed from equation 4.2 is 1.178. The estimated variance of $\hat{\mu}_Y$ computed from equation 4.7 is:

$$\text{Var}(\hat{\mu}_Y) = \text{EXP}[a_0 + a_1/(\mu_X+7)](-a_1(\mu_X+7)^{-2})\sigma_Y^2 + \sigma_X^2 + 2\text{EXP}[a_0 + a_1/(\mu_X+7)](-a_1(\mu_X+7)^{-2})\sigma_X\sigma_Y \rho_{XY} \quad [4.17]$$
For the deer data presented in Table 2.2, this estimate is .1555. The 95% confidence interval is (.405, 1.951).

Again, with the assumptions that Graybill (1977) requires, the exact calculation of $\text{Var}(\ln(\hat{\mu}_y)) = .017768$ giving a 95% confidence interval for $\ln(\hat{\mu}_y)$ of (-.09735, .42518). If we exponentiate the limits of this confidence interval, we get a 95% confidence interval for $\mu_y$ of (.9072, 1.5299).

We obtain a smaller confidence interval with Graybill's calculation of $\text{Var}(\hat{\mu}_y)$ but the more conservative propagation of error technique calculation we have no assumptions on the variables or their distribution.
CHAPTER V
CONCLUSION

In estimating \( \mu_Y \) we have discussed several estimation procedures. For each procedure, the variance of \( \hat{\mu}_Y \) was also given.

The ratio estimator uses a no intercept model and estimated variance both described by Cochran (1977). We next tried the idea of a no intercept model and applied least square techniques to get estimates of \( \mu_Y \) and \( \text{Var}(\hat{\mu}_Y) \).

The simple linear regression estimator was next used. It gave considerably better results than the no intercept model or the ratio. Due to the non-linearity of the deer data (see Figure 2.2) we expanded the results to a general regression model.

The general regression model allows us the greatest flexibility in trying to fit the data. Here, we used two curvilinear models to fit the deer data. The results we obtained were that the model \( Y = \alpha_0 + \alpha_1/(X+7) + \epsilon \) did a little better as measured by the narrower confidence intervals than did the polynominal model.

The estimated variance of \( \hat{\mu}_Y \) for the general regression model is often difficult or impossible to compute. Graybill (1977) gives a method for computing \( \text{Var}(\hat{\mu}_Y) \) but we need to assume that the \( X \) values are fixed non-random variables or that the regression
of $Y$ on $X$ is linear.

The propagation of error technique that Deming (1948) suggests approximates the variance of $\beta_Y$ by using a truncated Taylor series expansion after substituting $f[E(X)]$ for $E[f(X)]$.

The computation of $\Sigma f(X_i)/n$ and $f(EX_i/n)$ for our sample will give us some idea of how close Deming's (1948) approximation is. This calculation turned out to be close for the models presented with the biggest difference being .246 for the model $Y = a_0 + a_1/(X+7) + \epsilon$.

The general propagation of error form for computing the estimated variance of $\beta_Y$ is given by equation 4.7. This method is a more easily computed estimate of the variance of $\beta_Y$ than the exact variance of $\beta_Y$ is.

When the propagation of error technique was applied to the simple linear regression model we obtained the same estimate of the variance of $\beta_Y$ that Cochran (1977) did. This was expected since a Taylor series for a linear function is the function.

Table 5.1 gives a summary of the confidence intervals and estimate of $\beta_Y$ for each model as applied to the deer data. The results in this table suggest that the simple linear regression model did the best. The ratio, no intercept, and simple linear results may be a little suspect because of the apparent curvilinear nature of the data.

The propagation of error technique gave a larger confidence interval with every model than did exact methods. If the
Table 5.1. Results obtained using the models and variance techniques discussed for the deer example.

<table>
<thead>
<tr>
<th>Model</th>
<th>( \hat{\beta}_Y )</th>
<th>Confidence Interval Width</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Cochran</td>
</tr>
<tr>
<td>Ratio</td>
<td>1.28</td>
<td>.782</td>
</tr>
<tr>
<td>Regression (no intercept)</td>
<td>.916</td>
<td></td>
</tr>
<tr>
<td>Simple Linear Regression</td>
<td>1.0757</td>
<td>.577</td>
</tr>
<tr>
<td>Polynominal</td>
<td>1.2468</td>
<td></td>
</tr>
<tr>
<td>( a_0 + a_1/(X+7) )</td>
<td>1.178</td>
<td>1.546</td>
</tr>
</tbody>
</table>
random nature of the $X$ variable is accounted for and then approximated, these conservative results are understood.

The purpose of using an additional $X$ variable is that because of its correlation with $Y$ we are able to increase our precision in estimating $\mu_Y$. If we ignore the $X$ variable and estimate $\mu_Y$ using just the $Y$ variables we get $\hat{\mu}_Y = \bar{Y} = 1.002$ with a 95% confidence interval of $(.488, 1.515)$.

From this we can see that the propagation of error technique did not increase the precision of this estimator as measured by the confidence interval width. Using Graybill's (1977) results though, we obtained the desired shorter confidence intervals.

This difference between the confidence intervals does seem to be larger than expected, suggesting that perhaps the approximation is too crude. Taking further terms in the Taylor series might improve the approximation. Further work should be considered in finding out when this approximation is valid or what restrictions or assumptions need to be applied.

The results presented in this paper should not be considered totally definitive or exhaustive. Other techniques which might be used to get estimates of the variance of $\hat{\mu}_Y$ include the jackknife technique, repeated samples or subsampling. Monte Carlo studies would be necessary to determine which, if any, of these methods are best.

Perhaps the biggest problem with these methods is the assumption required that we must know the $\mu_X$'s. When this is
Perhaps the biggest problem with these methods is the assumption required that we must know the $\mu_X$'s. When this is combined with the bias introduced by Jensen's inequality, where we replace $E[f(X)]$ with $f[E(X)]$, the method may be seen to have serious limitations.
REFERENCES


