5.5 Regression Estimation

- Assume a SRS of n pairs $(x_1, y_1), \ldots, (x_n, y_n)$ is selected from a population of N pairs of (x, y) data. The goal of **regression estimation** is to take advantage of a linear relationship between x and y to improve estimation of the t_y or $\overline{y}_U$.
- Unlike ratio estimation, there is no assumption of a zero intercept in the linear relationship or a positive slope. We assume a linear form for the relationship between y and x:

$$y_i \;=\; B_0 \;+\; B_1 x_i \;+\; \epsilon_i \tag{53}$$

  for intercept B_0, slope B_1, and ϵ_i is the deviation between y_i and $B_0 + B_1 x_i$.
- We assume that (i) the mean of the ϵ_i's is zero and (ii) are uncorrelated with the x_i's. This implies that there is no systematic relationship between the e_i's and x_i's.
- For example, we do not want the variance to increase (or decrease) with the mean.

5.5.1 Estimating $\overline{y}_U$ and t_y

- To estimate $\overline{y}_U$, we first must get estimates $\widehat{B}_0$ and $\widehat{B}_1$ of the true intercept B_0 and slope B_1. We use the least squares estimates

$$\widehat{B}_0 \;=\; \overline{y} - \widehat{B}_1 \overline{x} \qquad\qquad \widehat{B}_1 \;=\; \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} \;=\; \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

- When $\overline{x}_U$ is <u>known</u>, our estimate $\widehat{\overline{y}}_{reg}$ of the population mean is:

$$\widehat{\overline{y}}_{reg} \;=\; \widehat{B}_0 \;+\; \widehat{B}_1 \overline{x}_U \tag{54}$$

  When $\overline{x}_U$ is <u>unknown</u>, replace $\overline{x}_U$ with $\overline{x}$ in (54). Then the estimate is $\widehat{\overline{y}}_{reg} = \overline{y}$.
- The estimated variance of $\widehat{\overline{y}}_{reg}$ in (54) is

$$\widehat{V}(\widehat{\overline{y}}_{reg}) \;=\; \frac{N-n}{N} \, \frac{1}{n(n-2)} \sum_{i=1}^n \tag{55}$$

- If N is unknown, but N is large relative to n (or, n/N is small), then the f.p.c. $(N-n)/N \approx 1$. Some researchers will replace $(N-n)/N$ with 1 in the variance formula in (55):

$$\widehat{V}(\widehat{\overline{y}}_{reg}) \;\approx\; \sum_{i=1}^n (y_i - \widehat{B}_0 - \widehat{B}_1 x_i)^2 \tag{56}$$

- An alternative formula for calculation:

$$\sum_{i=1}^n (y_i - \widehat{B}_0 - \widehat{B}_1 x_i)^2 \;=\; \sum_{i=1}^n y_i^2 \;-\; n\overline{y}^2 \;-\; \widehat{B}_1^2 \left(\sum_{i=1}^n x_i^2 \;-\; n\overline{x}^2 \right) \tag{57}$$

  which can be substituted into (55) to get the estimated variance $\widehat{V}(\widehat{\overline{y}}_{reg})$

$$\widehat{V}(\widehat{\overline{y}}_{reg}) \;=\; \frac{N-n}{Nn(n-2)} \left[\sum_{i=1}^n y_i^2 \;-\; n\overline{y}^2 \;-\; \widehat{B}_1^2 \left(\sum_{i=1}^n x_i^2 \;-\; n\overline{x}^2 \right) \right] \tag{58}$$

- An approximate $100(1-\alpha)$ confidence interval for $\overline{y}_U$ is $\widehat{\overline{y}}_{reg} \pm t^* \sqrt{\widehat{V}(\widehat{\overline{y}}_{reg})}$ where t^* is the upper $\alpha/2$ critical value from a t-distribution having $n-2$ degrees of freedom.
- By multiplying $\widehat{\overline{y}}_{reg}$ in (54) by N, an estimator $\widehat{t}_{reg}$ of the population total t_y is:

$$\begin{aligned} \widehat{t}_{reg} &= N\widehat{\overline{y}}_{reg} = N(\widehat{B}_0 + \widehat{B}_1\overline{x}_U) = N(\overline{y} - \widehat{B}_1\overline{x} + \widehat{B}_1\overline{x}_U) \\ &= N\overline{y} + \widehat{B}_1(N\overline{x}_U - N\overline{x}) = N\overline{y} + \widehat{B}_1(t_x - N\overline{x}) \end{aligned} \tag{59}$$

- Multiplying $\widehat{V}(\widehat{\overline{y}}_{reg})$ in (55) by N^2 provides the estimated variance of $\widehat{t}_{reg}$:

$$\widehat{V}(\widehat{t}_{reg}) = \frac{N(N-n)}{n(n-2)} \sum_{i=1}^{n} (y_i - \widehat{B}_0 - \widehat{B}_1 x_i)^2. \tag{60}$$

- An approximate $100(1-\alpha)$ confidence interval for t_y is $\widehat{t}_{reg} \pm t^* \sqrt{\widehat{V}(\widehat{t}_{reg})}$ where t^* is the upper $\alpha/2$ critical value from a t-distribution having $n-2$ degrees of freedom.
- Note that $\sum_{i=1}^{n} (y_i - \widehat{B}_0 - \widehat{B}_1 x_i)^2$ is the sum of squared residuals from a simple linear regression. That is, $\sum_{i=1}^{n} (y_i - \widehat{B}_0 - \widehat{B}_1 x_i)^2 = SSE$ for the least squares regression line.
- Therefore, we could fit a regression model using a statistics package, and substitute the value of SSE or MSE into (55) and get

$$\widehat{V}(\widehat{\overline{y}}_{reg}) =$$

$$\widehat{V}(\widehat{t}_{reg}) = \tag{61}$$

Example: The Florida Game and Freshwater Fish Commission is interested in estimating the weights of alligators using length measurements which are easier to observe. The population size N is unknown but is large enough so that ignoring the finite population correction will have a negligible effect on estimation. That is, $\frac{N-n}{N} \approx 1$ (see Equation (56)). A random sample of 22 alligators yielded the following weight (in pounds) and length (in inches) data:

Alligator	Length	Weight	Alligator	Length	Weight
1	94	130	12	86	83
2	74	51	13	88	70
3	82	80	14	72	61
4	58	28	15	74	54
5	86	80	16	61	44
6	94	110	17	90	106
7	63	33	18	89	84
8	86	90	19	68	39
9	69	36	20	76	42
10	72	38	21	78	57
11	85	84	22	90	102

Summary values for $n = 22$			
$\overline{x} = 78.8636364$	$\sum x_i = 1735$	$\sum x_i^2 = 139293$	
$\overline{y} = 68.2727273$	$\sum y_i = 1502$	$\sum y_i^2 = 119762$	$\sum x_i y_i = 124433$

- Because it is much easier to collect data on alligator length, there is a lot of available data on length. Assume that the available data indicates that the mean alligator length $\overline{x}_U \approx 90$ inches. Estimate the mean alligator weight $\overline{y}_U$ using regression estimator $\widehat{\overline{y}}_{reg}$.

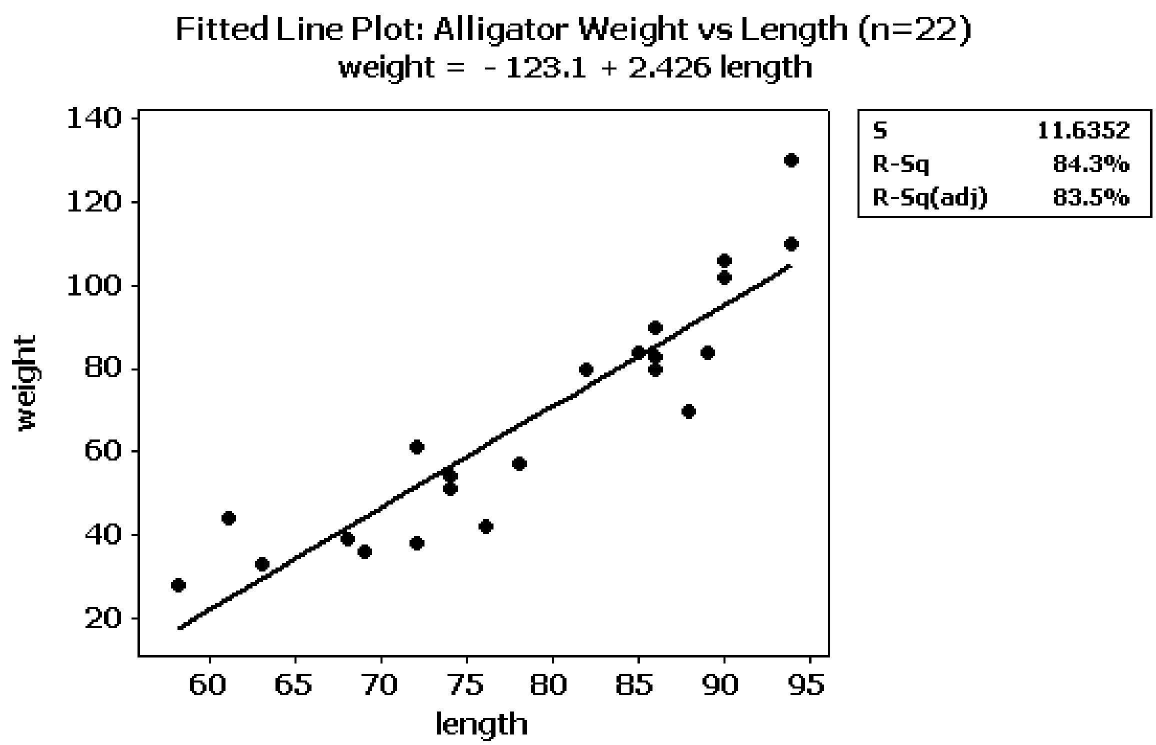
Regression output from MINITAB for the regression of alligator weight vs length ($n = 22$).

```
The regression equation is
weight = -123.074 + 2.42629 length

S = 11.6352      R-Sq = 84.3 %      R-Sq(adj) = 83.5 %

Analysis of Variance

Source             DF          SS          MS         F       P
Regression          1     14508.8     14508.8   107.172   0.000
Error              20      2707.6       135.4
Total              21     17216.4
```



- Suppose the sample size was actually $n = 25$ and included the following 3 measurements. Estimate the mean alligator weight $\overline{y}_U$ using the regression estimator $\widehat{\overline{y}}_{reg}$.

Alligator	Length	Weight
23	147	640
24	128	366
25	114	197

Summary values for $n = 25$

$\overline{x} = 84.96$	$\sum x_i = 2124$	$\sum x_i^2 = 190282$	
$\overline{y} = 108.2$	$\sum y_i = 2705$	$\sum y_i^2 = 702127$	$\sum x_i y_i = 287819$

- Do you have any concerns about using a regression estimator for this data? If so, how can we adjust our analysis to account for it?

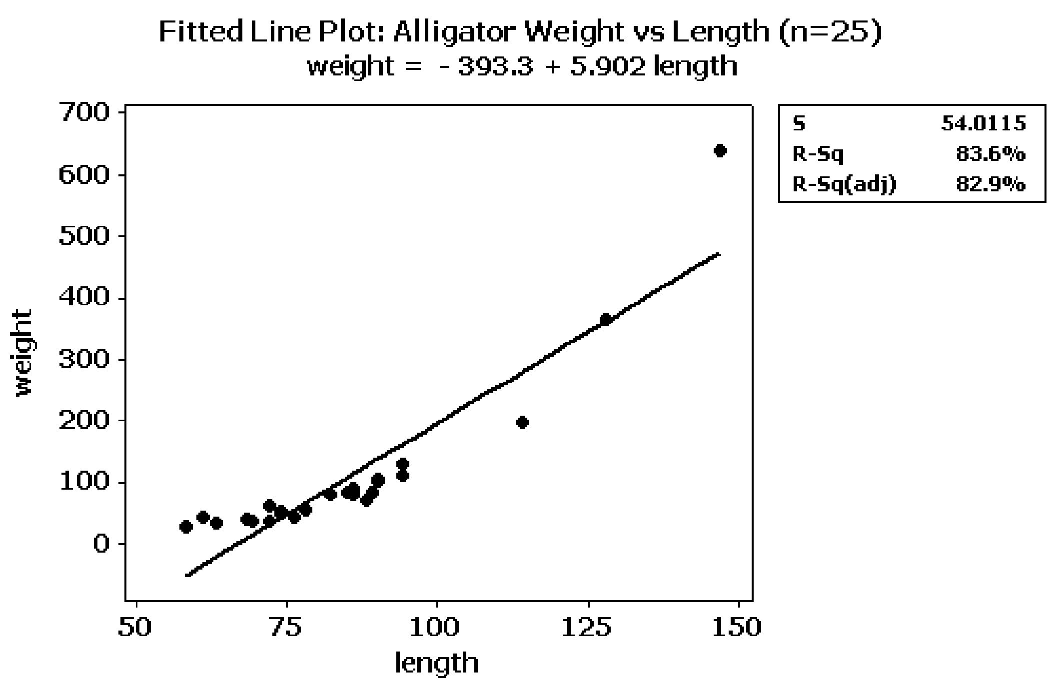
Regression output from MINITAB for the regression of alligator weight vs length ($n = 25$).

```
The regression equation is
weight = -393.264 + 5.90235 length

S = 54.0115      R-Sq = 83.6 %      R-Sq(adj) = 82.9 %

Analysis of Variance

Source            DF         SS          MS         F       P
Regression         1     342350      342350   117.354  0.000
Error             23      67096        2917
Total             24     409446
```



5.5.2 Extension to Multiple Regression

We can generalize the simple linear regression model to a multiple linear regression model with

Case 1: k different regression variables $x_1, x_2, \ldots, x_k$ with model

$$y \ = \ B_0 \ + \tag{62}$$

Case 2: A k^{th}-order polynomial in 1 regression variable x with model

$$y \ = \ B_0 \ + \ B_1 x \ + \ B_2 x^2 \ + \cdots + \ B_k x^k \ + \ \epsilon \ \ = \tag{63}$$

For Case 1:

1. Find the least-squares estimates $(\widehat{B}_0, \widehat{B}_1, \ldots, \widehat{B}_k)$. This will produce the prediction model:

$$\widehat{y} \ = \ \widehat{B}_0 \ + \tag{64}$$

2. To estimate $\overline{y}_U$, replace each variable with its mean:

$$\widehat{\overline{y}}_{reg} \ = \ \widehat{B}_0 \ + \tag{65}$$

where $\overline{x}_{Ui}$ is the mean of x_i.

For Case 2:

1. Find the least-squares estimates $(\widehat{B}_0, \widehat{B}_1, \ldots, \widehat{B}_k)$. This will produce the prediction model:

$$\widehat{y} \ = \ \widehat{B}_0 \ + \ \widehat{B}_1 x \ + \ \widehat{B}_2 x^2 \ + \cdots + \ \widehat{B}_k x^k \tag{66}$$

2. To estimate $\overline{y}_U$, replace x with its mean:

$$\widehat{\overline{y}}_{reg} \ = \ \widehat{B}_0 \ + \ \widehat{B}_1 \overline{x}_U \ + \ \widehat{B}_2 \overline{x}_U^2 \ + \cdots + \ \widehat{B}_k \overline{x}_U^k \tag{67}$$

- For Case 1 and for Case 2, we use the SSE from the regression output to calculate the estimated variance of $\widehat{\overline{y}}_{reg}$:

$$\widehat{V}(\widehat{\overline{y}_{reg}}) \ = \ \frac{N-n}{Nn(n-k-1)} SSE \ = \tag{68}$$

where $MSE = SSE/(n-k-1)$ is the mean squared error from the regression having $n-k-1$ degrees of freedom for the Error term.

Example: Use the data from the Florida Game and Freshwater Fish Commission example

- Fit the quadratic regression model $y = B_0 + B_1 x + B_2 x^2 + \epsilon$.
- Estimate the mean alligator weight $\overline{y}_U$ using the multiple regression estimator $\widehat{\overline{y}}_{reg}$.
- Find a 95% confidence interval for $\overline{y}_U$.

Regression output from MINITAB for the quadratic regression of alligator weight vs length ($n = 25$).

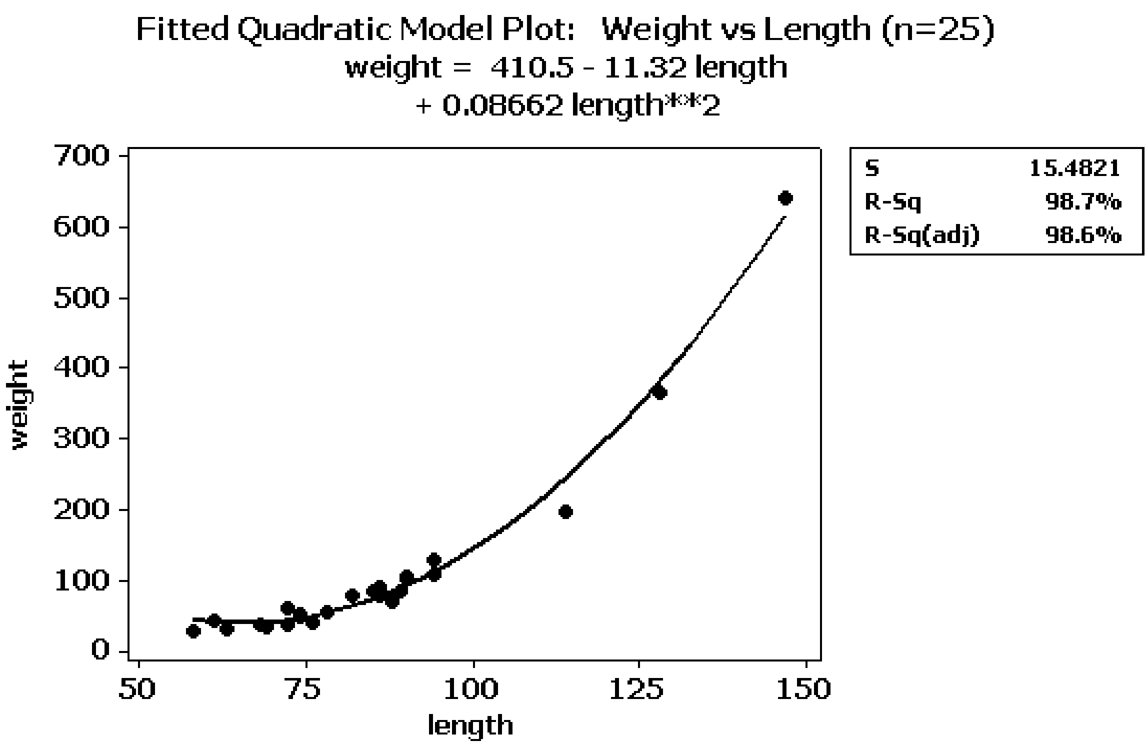
```
The regression equation is
weight = 410.484 - 11.3176 length + 0.0866155 length**2

S = 15.4821      R-Sq = 98.7 %      R-Sq(adj) = 98.6 %

Analysis of Variance

Source            DF        SS        MS        F      P
Regression         2    404173    202086  843.102  0.000
Error             22      5273       240
Total             24    409446

Source     DF    Seq SS        F      P
Linear      1    342350  117.354  0.000
Quadratic   1     61823  257.926  0.000
```



5.6 Regression Analyses using R and SAS

5.6.1 Example with unknown N

R code for Regression Estimation (N unknown but large)

- **CASE 1**: We will fit the model gatorwgt $=$ $\widehat{B}_0$ $+$ $\widehat{B}_1$ gatorlen using the original 22 pairs of alligator length (x) and weight (y) data. We will also estimate the mean weight $\overline{y}_U$ using regression estimation (assuming $\overline{x}_U = 90$ pounds). The df $= n - 2 = 20$ because the model has 2 parameters.
- Unfortunately, the 'survreg' function in R uses $Nn(n-1)$ in the estimated variance formula (see equation (54) instead of $Nn(n-2)$ for linear regression.
- If the regression model has p parameters, we want $Nn(n-p)$ in the variance formula.
- Thus, if we multiply the standard error given in R by $\sqrt{(n-1)/(n-p)} = \sqrt{(n-1)/df}$ where $df = n - p$ the regression degrees of freedom we get the same analysis as SAS.

R code for Regression Estimation: Case 1

```
library(survey)

# Regression estimation

x <- c(94,74,82,58,86,94,63,86,69,72,85,86,88,72,74,61,90,89,68,76,78,90)
y <- c(130,51,80,28,80,110,33,90,36,38,84,83,70,61,54,44,106,84,39,42,57,102)

ci.level = .95   # confidence level                          <---
p = 2            # number of parameters in the model         <---
n = length(x)
df = n - p
corr.fctr = sqrt((n-1)/df)

regdata <- data.frame(x,y)                                   <---

regdsgn <- svydesign(id=~1,data=regdata)
regdsgn

svyreg <- svyglm(y~x,design=regdsgn)              <-- enter regression model
svyreg
atmean <- c(1,90)                                 <-- enter x mean = 90

# Assign names to each term in the model

names(atmean) <- c("(Intercept)","x")             <-- two model terms
atmean <- as.data.frame(t(atmean))

meanwgt <- predict(svyreg,newdata=atmean,total=1)
meanwgt

# Enter standard error
se.meanwgt = 4.0207                           <-- value output by meanwgt

# Compute confidence interval for mean with correction factor
meanwgt + c(-1,1)*corr.fctr*qt(ci.level+(1-ci.level)/2,df)*se.meanwgt
```

R output for Regression Estimation: Case 1

```
Independent Sampling design (with replacement)
svydesign(id = ~1, data = regdata)

Coefficients:
(Intercept)            x
   -123.074        2.426

Degrees of Freedom: 21 Total (i.e. Null);  20 Residual
Null Deviance:      17220
Residual Deviance: 2708         AIC: 174.3

> meanwgt
    link     SE
1 95.293 4.0207
>
> # Enter standard error
> se.meanwgt = 4.0207
>
> # Compute confidence interval for mean with correction factor
[1]  86.69865 103.88695
```

- **CASE 2**: We will fit the model gatorwgt $= \widehat{B}_0 + \widehat{B}_1$ gatorlen using the 25 pairs of alligator length (x) and weight (y) data.
- We will again estimate the mean weight $\overline{y}_U$ using regression estimation (assuming $\overline{x}_U = 90$ pounds).

R code for Regression Estimation: Case 2

```
library(survey)

# Regression estimation

x <- c(94,74,82,58,86,94,63,86,69,72,85,86,88,72,74,61,90,89,68,76,78,90,
   147,128,114)
y <- c(130,51,80,28,80,110,33,90,36,38,84,83,70,61,54,44,106,84,39,42,57,102,
   640,366,197)

ci.level = .95   # confidence level                           <--
p = 2            # number of parameters in the model          <--
n = length(x)
df = n - p
corr.fctr = sqrt((n-1)/df)

regdata <- data.frame(x,y)                          <--

regdsgn <- svydesign(id=~1,data=regdata)
regdsgn

svyreg <- svyglm(y~x,design=regdsgn)    <-- enter regression model
svyreg
atmean <- c(1,90)                                   <-- enter x mean = 90
```

```
# Assign names to each term in the model

names(atmean) <- c("(Intercept)","x")          <-- two model terms
atmean <- as.data.frame(t(atmean))

meanwgt <- predict(svyreg,newdata=atmean,total=1)
meanwgt

# Enter standard error
se.meanwgt = 14.542                                <-- value output by meanwgt

# Compute confidence interval for mean with correction factor
meanwgt + c(-1,1)*corr.fctr*qt(ci.level+(1-ci.level)/2,df)*se.meanwgt
```

R output for Regression Estimation: Case 2

```
Independent Sampling design (with replacement)

Coefficients:
(Intercept)            x
   -393.264        5.902

Degrees of Freedom: 24 Total (i.e. Null);  23 Residual
Null Deviance:      409400
Residual Deviance: 67100        AIC: 274.3

> meanwgt
    link     SE
1 137.95 14.542

> # Enter standard error
> se.meanwgt = 14.542

> # Compute confidence interval for mean with correction factor
[1] 107.2184 168.6773
```

- **CASE 3**: We will fit the model

$$\text{gatorwgt} \; = \; \widehat{B}_0 \; + \; \widehat{B}_1 \, \text{gatorlen} \; + \; \widehat{B}_2 \, \text{gatorlen}^2$$

  using the original 22 pairs of alligator length (x) and weight (y) data.

- We will estimate the mean weight $\overline{y}_U$ using quadratic model regression estimation (assuming $\overline{x}_U = 90$ pounds). The df $= n - 3 = 19$ because the model has 3 parameters.

R code for regression estimation: Case 3

```
library(survey)

# Regression estimation

x <- c(94,74,82,58,86,94,63,86,69,72,85,86,88,72,74,61,90,89,68,76,78,90)
y <- c(130,51,80,28,80,110,33,90,36,38,84,83,70,61,54,44,106,84,39,42,57,102)
```

```
xsq <- x^2                <-- create x-squared values for quadratic model
xsq

ci.level = .95   # confidence level                   <--
p = 3            # number of parameters in the model  <--
n = length(x)
df = n - p
corr.fctr = sqrt((n-1)/df)

regdata <- data.frame(x,xsq,y)                        <--

regdsgn <- svydesign(id=~1,data=regdata)
regdsgn

svyreg <- svyglm(y~x+xsq,design=regdsgn)              <--
svyreg
atmean <- c(1,90,8100)                                <--

# Assign names to each term in the model

names(atmean) <- c("(Intercept)","x","xsq")           <--
atmean <- as.data.frame(t(atmean))

meanwgt <- predict(svyreg,newdata=atmean,total=1)
meanwgt

# Enter standard error
se.meanwgt = 3.3693                                   <--

# Compute confidence interval for mean with correction factor
meanwgt + c(-1,1)*corr.fctr*qt(ci.level+(1-ci.level)/2,df)*se.meanwgt
```

R output for regression estimation: Case 3

```
Independent Sampling design (with replacement)

Coefficients:
(Intercept)            x          xsq
  269.71588     -7.97504      0.06752

Degrees of Freedom: 21 Total (i.e. Null);  19 Residual
Null Deviance:      17220
Residual Deviance: 1628         AIC: 165.1

> meanwgt
    link     SE
1 98.867 3.3693

> # Enter standard error
> se.meanwgt = 3.3693

> # Compute confidence interval for mean with correction factor
[1]  91.4535 106.2813
```

- **CASE 4**: We will fit the model

$$\text{gatorwgt} = \widehat{B}_0 + \widehat{B}_1\,\text{gatorlen} + \widehat{B}_2\,\text{gatorlen}^2$$

  using the 25 pairs of alligator length (x) and weight (y) data.

- We will again estimate the mean weight $\overline{y}_U$ using quadratic model regression estimation (assuming $\overline{x}_U = 90$ pounds).

R code for regression estimation: Case 4

```
library(survey)

# Regression estimation

x <- c(94,74,82,58,86,94,63,86,69,72,85,86,88,72,74,61,90,89,68,76,78,90,
147,128,114)
y <- c(130,51,80,28,80,110,33,90,36,38,84,83,70,61,54,44,106,84,39,42,57,102,
640,366,197)

xsq <- x^2                  <-- create x-squared values for quadratic model
xsq

ci.level = .95   # confidence level                       <--
p = 3            # number of parameters in the model      <--
n = length(x)
df = n - p
corr.fctr = sqrt((n-1)/df)

regdata <- data.frame(x,xsq,y)                            <--

regdsgn <- svydesign(id=~1,data=regdata)
regdsgn

svyreg <- svyglm(y~x+xsq,design=regdsgn)                  <--
svyreg
atmean <- c(1,90,8100)                                    <--

# Assign names to each term in the model

names(atmean) <- c("(Intercept)","x","xsq")               <--
atmean <- as.data.frame(t(atmean))

meanwgt <- predict(svyreg,newdata=atmean,total=1)
meanwgt

# Enter standard error
se.meanwgt = 4.9589                                       <--

# Compute confidence interval for mean with correction factor
meanwgt + c(-1,1)*corr.fctr*qt(ci.level+(1-ci.level)/2,df)*se.meanwgt
```

R output for regression estimation: Case 4

```
Independent Sampling design (with replacement)
svydesign(id = ~1, data = regdata)

Coefficients:
(Intercept)            x          xsq
  410.48412    -11.31755      0.08662

Degrees of Freedom: 24 Total (i.e. Null);  22 Residual
Null Deviance:      409400
Residual Deviance: 5273         AIC: 212.7

> meanwgt
   link     SE
1 93.49 4.9589

> # Enter standard error
> se.meanwgt = 4.9589

> # Compute confidence interval for mean with correction factor
[1]  82.74886 104.23171
```

SAS code for Regression Estimation (N unknown but large)

We will reproduce the R analyses for Case 1 (linear regression with $n = 22$ points) and Case 4 (quadratic regression with $n = 25$ points). For Case 2 and Case 4, you just have to change the data set.

- **CASE 1**: We will fit the model

$$\text{gatorwgt} \;=\; \widehat{B}_0 \;+\; \widehat{B}_1\,\text{gatorlen}$$

  using the original 22 pairs of alligator length (x) and weight (y) data. We will also estimate the mean weight $\overline{y}_U$ using regression estimation (assuming $\overline{x}_U = 90$ pounds). The df $=$ $n - 2 = 20$ because the model has 2 parameters.

```
data in; input gatorid gatorlen gatorwgt @@;
cards;
1   94 130     2  74  51     3  82  80     4  58  28     5  86  80
6   94 110     7  63  33     8  86  90     9  69  36    10  72  38
11  85  84    12  86  83    13  88  70    14  72  61    15  74  54
16  61  44    17  90 106    18  89  84    19  68  39    20  76  42
21  78  57    22  90 102
;
/* Use proc surveyreg to estimate the average alligator weight */

proc surveyreg data=in ;
    model gatorwgt = gatorlen / clparm solution df=20;

/* To estimate a mean substitute 1 for intercept, mean(x) for gatorwgt */
   estimate 'Average gator weight' intercept 1 gatorlen 90;
run;
```

SAS output for Case 1

```
                         The SURVEYREG Procedure

           Regression Analysis for Dependent Variable gatorwgt

                              Data Summary

                   Number of Observations           22
                   Mean of gatorwgt           68.27273
                   Sum of gatorwgt              1502.0

                             Fit Statistics

                        R-square            0.8427
                        Root MSE           11.6352
                        Denominator DF          20

                         Tests of Model Effects

                 Effect       Num DF    F Value    Pr > F

                 Model             1      80.32    <.0001
                 Intercept         1      33.68    <.0001
                 gatorlen          1      80.32    <.0001

      NOTE: The denominator degrees of freedom for the F tests is 20.

                     Estimated Regression Coefficients

                         Standard                          95% Confidence
Parameter    Estimate       Error  t Value  Pr > |t|          Interval

Intercept  -123.07351  21.2067045    -5.80    <.0001  -167.30992 -78.837103
gatorlen      2.42629   0.2707230     8.96    <.0001     1.86157   2.991011

                      Analysis of Estimable Functions

                                             Standard
  Parameter                 Estimate            Error    t Value    Pr > |t|

  Average gator weight       95.2928          4.11999      23.13      <.0001

                      Analysis of Estimable Functions

                                              95% Confidence
             Parameter                           Interval

             Average gator weight       86.6987    103.887
      NOTE: The denominator degrees of freedom for the t tests is 20.
```

- **CASE 4**: We will fit the model $\text{gatorwgt} = \widehat{B}_0 + \widehat{B}_1 \,\text{gatorlen} + \widehat{B}_2 \,\text{gatorlen}^2$ using the 25 pairs of alligator length (x) and weight (y) data. We will again estimate the mean weight $\overline{y}_U$ using quadratic model regression estimation (assuming $\overline{x}_U = 90$ pounds).

```
This is the data for Case 4. Change df=19 to df=22 in the SAS code.

data in; input gatorid gatorlen gatorwgt @@;
     gatorsq = gatorlen**2;
cards;
1   94 130     2  74  51     3  82  80     4  58  28     5  86  80
6   94 110     7  63  33     8  86  90     9  69  36    10  72  38
11  85  84    12  86  83    13  88  70    14  72  61    15  74  54
16  61  44    17  90 106    18  89  84    19  68  39    20  76  42
21  78  57    22  90 102    23  147 640    24 128 366    25 114 197
;
```

SAS output for Case 4

```
                          The SURVEYREG Procedure

            Regression Analysis for Dependent Variable gatorwgt

                              Data Summary

                    Number of Observations           25
                    Mean of gatorwgt          108.20000
                    Sum of gatorwgt              2705.0

                             Fit Statistics

                         R-square           0.9871
                         Root MSE          15.4821
                         Denominator DF         22

                         Tests of Model Effects

                  Effect       Num DF    F Value    Pr > F

                  Model             2     389.43    <.0001
                  Intercept         1      47.97    <.0001
                  gatorlen          1      81.42    <.0001
                  gatorsq           1     188.80    <.0001

      NOTE: The denominator degrees of freedom for the F tests is 22.

                     Estimated Regression Coefficients

                         Standard                            95% Confidence
Parameter    Estimate       Error  t Value  Pr > |t|            Interval

Intercept  410.484123  59.2682925     6.93    <.0001  287.569207 533.399038
gatorlen   -11.317553   1.2542503    -9.02    <.0001  -13.918708  -8.716397
gatorsq      0.086616   0.0063037    13.74    <.0001    0.073542   0.099689

      NOTE: The denominator degrees of freedom for the t tests is 22.

                      Analysis of Estimable Functions

                                                Standard
  Parameter                    Estimate            Error    t Value    Pr > |t|
  Average gator weight       93.4902828       5.17936779      18.05      <.0001

                      Analysis of Estimable Functions

                                                 95% Confidence
             Parameter                              Interval
             Average gator weight    82.7489314 104.231634

      NOTE: The denominator degrees of freedom for the t tests is 22.
```

5.6.2 Example with known N and t_x

The data is from Lohr Example 4.9 (page 139-141): To estimate the number of dead trees in a study area, we divide the study area into 100 square plots and count the number of dead trees in a photograph of each plot. Photo counts can be made quickly, but sometimes a tree is misclassified or not detected. A SRS of 25 of the plots for field counts of dead trees (ground truthing) is taken. The population mean number of dead trees per plot from the photo counts is 11.3. The data is given in the R code below. Estimate the actual mean number of dead trees per plot and the total number of dead trees in the study area.

R code for regression estimation – t_x and N known

```
library(survey)

# Regression estimation with known tx and N (Lohr Example 4.9)

x <- c(10,12,7,13,13,6,17,16,15,10,14,12,10,5,12,10,10,9,6,11,7,9,11,10,10)
y <- c(15,14,9,14,8,5,18,15,13,15,11,15,12,8,13,9,11,12,9,12,13,11,10,9,8)

ci.level = .95   # confidence level                         <---
p = 2            # number of parameters in the model        <---
n = length(x)
df = n - p
corr.fctr = sqrt((n-1)/df)

fpc <- c(rep(100,n))                                   <--- 100 = N
regdata <- data.frame(x,y)

regdsgn <- svydesign(id=~1,fpc=~fpc,data=regdata)      <--- include fpc
regdsgn

svyreg <- svyglm(y~x,design=regdsgn)
svyreg
atmean <- c(1,11.3)                               <-- evaluate at x-mean

# Assign names to each term in the model

names(atmean) <- c("(Intercept)","x")
atmean <- as.data.frame(t(atmean))

meanwgt <- predict(svyreg,newdata=atmean,total=1)
meanwgt

# Enter standard error
se.meanwgt = .418                            <-- input s.e. from R

# Compute confidence interval for mean with correction factor
meanwgt + c(-1,1)*corr.fctr*qt(ci.level+(1-ci.level)/2,df)*se.meanwgt

# Compute confidence interval for a total with correction factor
ttl = 100     # ttl is the total number of units
sumwgt <- svycontrast(meanwgt,ttl)
sumwgt + c(-1,1)*corr.fctr*qt(ci.level+(1-ci.level)/2,df)*se.meanwgt*ttl
```

R output for regression estimation – t_x and N known

```
Coefficients:
(Intercept)            x
     5.0593       0.6133

Degrees of Freedom: 24 Total (i.e. Null);  23 Residual
Null Deviance:      218.2
Residual Deviance: 133.2        AIC: 118.8

> meanwgt
    link    SE
1 11.989 0.418

> # Enter standard error
> se.meanwgt = .418

> # Compute confidence interval for mean with correction factor
[1] 11.10600 12.87259

> # Compute confidence interval for a total with correction factor
[1] 1110.600 1287.259
```

SAS code for regression estimation – t_x and N known (supplemental)

```
DATA trees;
 INPUT photo field @@;
   N = 100;
DATALINES;
10 15   12 14    7  9   13 14   13  8    6  5   17 18
16 15   15 13   10 15   14 11   12 15   10 12    5  8
12 13   10  9   10 11    9 12    6  9   11 12    7 13
 9 11   11 10   10  9   10  8
;
/* Use proc surveyreg to estimate the total number of trees */

PROC SURVEYREG DATA=trees TOTAL=100;
     MODEL field = photo / clparm solution df=23;
     ESTIMATE 'Total field trees' intercept 100 photo 1130;

/* substitute N for intercept, t_x for photo */
   ESTIMATE 'Mean field trees' intercept 100 photo 1130 / divisor=100;
run;
```

SAS output for regression estimation – t_x and N known

The SURVEYREG Procedure

Regression Analysis for Dependent Variable field

Data Summary

Number of Observations	25
Sum of Weights	100.00000
Weighted Mean of field	11.56000
Weighted Sum of field	1156.0

Fit Statistics

R-square	0.3896
Root MSE	2.4062
Denominator DF	23

Tests of Model Effects

Effect	Num DF	F Value	Pr > F
Model	1	22.73	<.0001
Intercept	1	12.64	0.0017
photo	1	22.73	<.0001

NOTE: The denominator degrees of freedom for the F tests is 23.

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Interval	
Intercept	5.05929204	1.42291882	3.56	0.0017	2.11576019	8.00282388
photo	0.61327434	0.12863682	4.77	<.0001	0.34716880	0.87937987

NOTE: The denominator degrees of freedom for the t tests is 23.

Analysis of Estimable Functions

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Total field trees	1198.92920	42.7013825	28.08	<.0001
Mean field trees	11.98929	0.4270138	28.08	<.0001

Regression Analysis for Dependent Variable field

Analysis of Estimable Functions

Parameter	95% Confidence Interval	
Total field trees	1110.59466	1287.26374
Mean field trees	11.10595	12.87264

NOTE: The denominator degrees of freedom for the t tests is 23.

5.7 $\widehat{\overline{y}_U}$ vs $\widehat{\overline{y}}_{reg}$ or $\widehat{t}$ vs $\widehat{t}_{reg}$ Which is better? SRS or Regression Estimation?

- When does regression estimation provide better estimates of $\overline{y}_U$ or t_y than the simple random sample (SRS) estimator?
- Recall that for least squares regression we have $SS(total) = SS(regression) + SSE$. Then
  - If x and y are strongly correlated, most of the variability in the y values can be explained by the regression and there will be very little random variability about the regression line. That is, $MS(regression)$ is large relative to MSE.
  - If x and y are weakly correlated, very little of the variability in the y values can be explained by the regression and most of the variability is random variability about the regression line. That is, $MS(regresssion)$ is small relative to MSE.
- Thus, for a moderate to strong linear relationship between x and y, regression estimation is recommended over SRS estimation.

5.8 Sample Size Estimation for Ratio and Regression Estimation

- To determine the sample size formulas for ratio and regression estimation, we will use the same approach that was used for determining a sample size for simple random sampling:
  1. Specify a maximum allowable difference d for the parameter we want to estimate. This is equivalent to stating the largest margin of error the researcher would want for a confidence interval.
  2. Specify α (where $100(1-\alpha)\%$ is the confidence level for the confidence interval).
  3. Specify a prior estimate of a variance $V(\widehat{\theta})$ where $\widehat{\theta}$ is the estimator of parameter θ. θ could be $\overline{y}_U$ or t_y or population ratio $R = \overline{y}_U/\overline{x}_U$ (in ratio estimation).
  4. Set the margin of error formula equal to d, and solve for n.

Sample Size Determination for Ratio Estimation:

- From equations (50), (47), (48), and (49), the margin of error formulas for parameters B, $\overline{y}_U$, and t_y with estimated variance s_e^2 are

$$\begin{aligned}
\text{For } R: \quad z_{\alpha/2}\sqrt{\widehat{V}(r)} &= z_{\alpha/2}\sqrt{\left(\frac{N-n}{N\overline{x}_U^2}\right)\frac{s_e^2}{n}} \\
\text{For } \overline{y}_U: \quad z_{\alpha/2}\sqrt{\widehat{V}(\widehat{\mu}_r)} &= z_{\alpha/2}\sqrt{\left(\frac{N-n}{N}\right)\frac{s_e^2}{n}} \\
\text{For } t_y: \quad z_{\alpha/2}\sqrt{\widehat{V}(\widehat{t}_r)} &= z_{\alpha/2}\sqrt{N(N-n)\frac{s_e^2}{n}}
\end{aligned}$$

After setting the margin of error $= d$, and solving for n yields

For B : $n = \dfrac{1}{\frac{1}{n_0} + \frac{1}{N}}$ where $n_0 =$ where d is the maximum allowable difference for estimating the ratio B. It is assumed that you know $\overline{x}_U$. If you do not know $\overline{x}_U$, you must provide an estimate $\widehat{\overline{x}_U}$.

For $\overline{y}_U$: $n = \dfrac{1}{\frac{1}{n_0} + \frac{1}{N}}$ where $n_0 =$ where d is the maximum allowable difference for estimating the mean $\overline{y}_U$.

For t_y : $n = \dfrac{1}{\frac{1}{n_0} + \frac{1}{N}}$ where $n_0 =$ where d is the maximum allowable difference for estimating the total t_y.

- s_e^2 is an estimate of the variability of the (x, y) points about a line having a zero intercept. You can use s_e^2 from a prior study, a pilot study, or double sampling.
- **Example of sample size determination**: Using the information from the pulpwood and dry wood example, estimate the sample size required so that $\widehat{t_y}$ is within 1200 kg of t_y with a probability of .95. Assume the new truckloads contain 1000 bundles of wood.

Sample Size Estimation for Regression Estimation:

- From equation (61), (47) and the confidence interval formulas for $\overline{y}_U$ and t_y, the margin of error formulas for parameters $\overline{y}_U$, and t_y with estimated variance MSE are

$$\text{For } \overline{y}_U : \quad z_{\alpha/2}\sqrt{\widehat{V}(\widehat{\overline{y}}_{reg})} = z_{\alpha/2}\sqrt{\frac{N-n}{Nn}MSE}$$

$$\text{For } t_y : \quad z_{\alpha/2}\sqrt{\widehat{V}(\widehat{\tau}_{reg})} = z_{\alpha/2}\sqrt{\frac{N(N-n)}{n}MSE}$$

After setting the margin of error $= d$, and solving for n yields

For $\overline{y}_U$: $n = \dfrac{1}{\frac{1}{n_0} + \frac{1}{N}}$ where $n_0 =$ where d is the maximum allowable difference for estimating the mean $\overline{y}_U$.

For t_y : $n = \dfrac{1}{\frac{1}{n_0} + \frac{1}{N}}$ where $n_0 =$ where d is the maximum allowable difference for estimating the total t_y.

- The MSE for the regression is an estimate of the variability of the (x, y) points about the regression line. You can use MSE from a prior study, a pilot study, or double sampling.