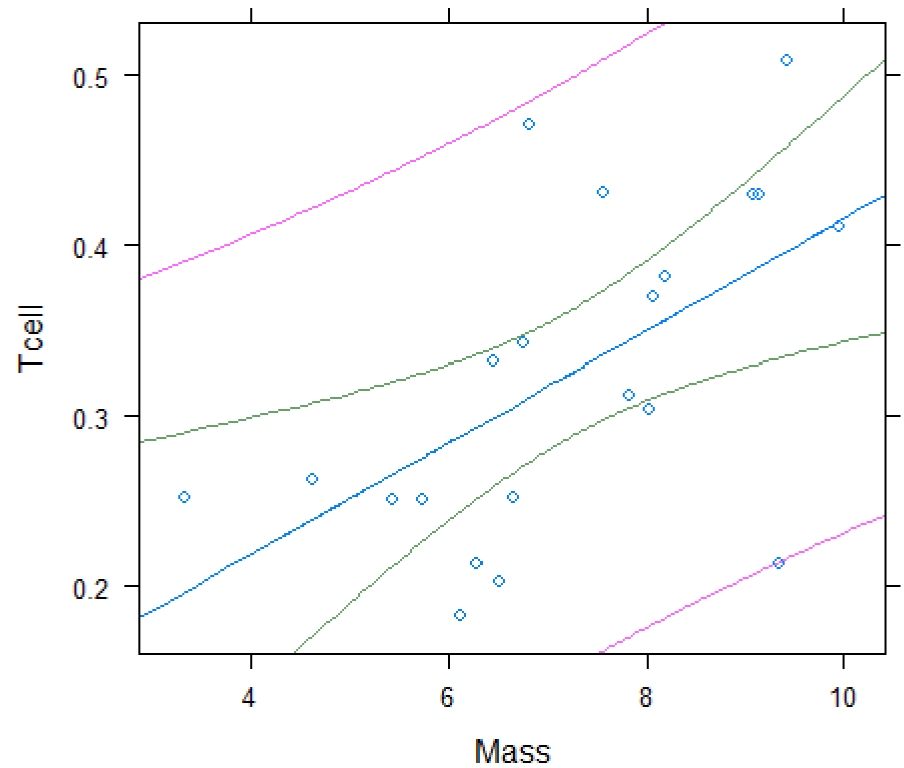
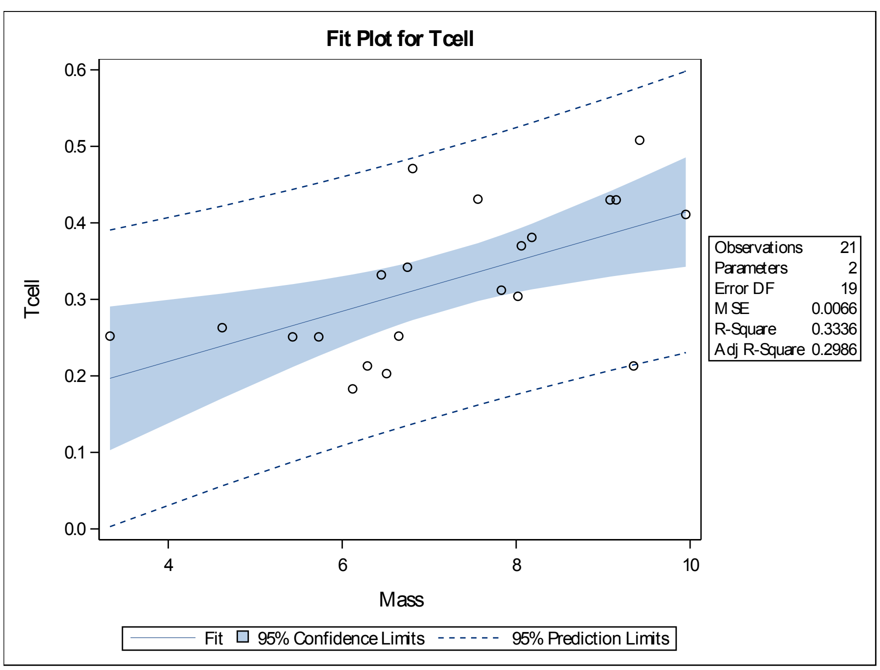
**Question 1)**

Part B)i) Scatterplots in R/SAS w/ Regression Confidence & Prediction Intervals



Part B)ii) Tabular Regression Parameter T-Stats and P-values in SAS/R

| Parameter Estimates | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | 95% Confidence Limits | |
| Intercept | 1 | 0.08750 | 0.07868 | 1.11 | 0.2800 | -0.07717 | 0.25217 |
| Mass | 1 | 0.03282 | 0.01064 | 3.08 | 0.0061 | 0.01055 | 0.05509 |

```
Coefficients:
            Estimate Std. Error t value  Pr(>|t|)

(Intercept)   0.0875     0.0787    1.11    0.2800
Mass          0.0328     0.0106    3.08    0.0061  **

---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

Part B)iii) Six-Step Hypothesis Tests for Regression Parameters

*Slope:*

$H_0$: $\beta_1 = 0$ $\quad T_{CRIT} = t_{.975,\,.025,\,19} = \pm\, 2.093024$

$H_A$: $\beta_1 \neq 0$ $\quad T_{STAT} = \frac{.03282 - 0}{.01064} = 3.084586$

$P_{VAL} = .006101$ → Reject $H_0$

→ There is **sufficient** evidence at the $\propto = .05$ level of significance ($P_{VAL} = .006101$) to suggest that: the value of the slope parameter coefficient is not equal to zero.

*Intercept*:

$H_0$: $\beta_0 = 0$ $\quad T_{CRIT} = t_{.975,\,.025,\,19} = \pm\, 2.093024$

$H_A$: $\beta_0 \neq 0$ $\quad T_{STAT} = \frac{.0875 - 0}{.07868} = 1.1121$

$P_{VAL} = .27997$ → FTR $H_0$

→ There is **not** sufficient evidence at the $\propto = .05$ level of significance ($P_{VAL} = .27997$) to suggest that: the value of the intercept parameter coefficient is not equal to zero.

Part B)iv) Regression Equation

*Model Estimate*: $\hat{y} = .0875 + .03282x$

Part B)v) Interpretation of Regression Parameters

*Parameter Interpretation*:

Slope: For every unit (g.) increase in Mass, the predicted T-Cell value (mm.) increases by .03282.

Intercept: The lack of statistical significance of the parameter estimate aside-- as there were no zero-valued observations of Mass (min: 3.33), the intercept is not of practical significance, though it could be interpreted, from the quantitative perspective of the regression estimate, as the point at which Mass no longer factors into the predicted T-cell value.

Part B)vi) & vii) 95% CI/PI & Interpretation: T-Cell Response at Stone Mass = 4.5 grams in R/SAS

*95% CI*: T-cell | Mass (4.5g) → [.1645, .3059] → (Fitted Value = .2352)

95% of the repetitions of the sampling process include a **mean** response associated with the above intervals at the specified explanatory value of Mass = 4.5g.

*95% PI*: T-cell | Mass (4.5g) → [.0515, .4189] → (Fitted Value = .2352)

95% of the repetitions of the sampling process include an **individual** response associated with the above intervals at the specified explanatory value of Mass = 4.5g.

```
> predict(lm_birds, interval = "confidence", data.frame(Mass = 4.5))

     fit    lwr    upr
1 0.2352 0.1645 0.3059

> predict(lm_birds, interval = "prediction", data.frame(Mass = 4.5))

     fit     lwr    upr
1 0.2352 0.05147 0.4189
```

**Output Statistics**

| Obs | Dependent Variable | Predicted Value | Std Error Mean Predict | 95% CL Mean | | 95% CL Predict | | Residual | Std Error Residual | Student Residual | Cook's D |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | . | 0.2352 | 0.0338 | 0.1645 | 0.3059 | 0.0515 | 0.4189 | . | . | . | . |

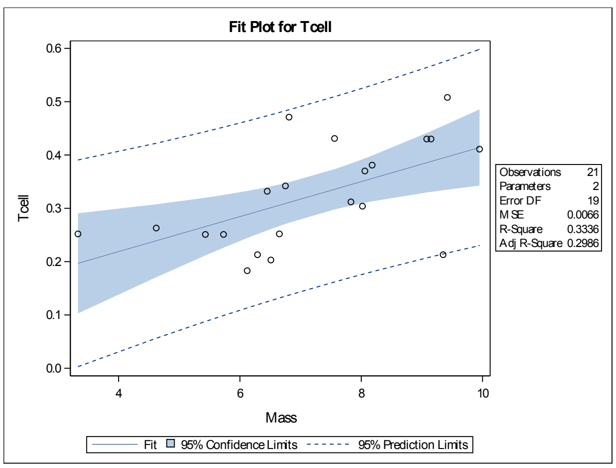
Part B)viii) Software/Graphical Method: Calibration CI/PI Intervals: T-cell Response = .3 in R/SAS

```
> calibrate(lm_birds, .3, interval = "Wald", mean.response = TRUE)

estimate    lower    upper       se
  6.4745   5.2431   7.7059   0.5883
```

```
> calibrate(lm_birds, .3, interval = "Wald", mean.response = FALSE)

estimate    lower    upper       se
   6.474    1.163   11.786    2.538
```

*Interpretation*:

At the specified mean/individual T-cell value, the respective 95% CI/PI for the corresponding measurement of Mass would be [5.2431, 7.7059] and [1.163, 11.786], respectively.




Extending a horizontal line outward from **Y = .3** and two vertical lines down from its points of intersection with the indicated CI and PI bands, one can estimate the **inverse 95% CI {X | Y$_0$}** to be approximately [5.1, 7.8] and the **inverse 95% PI {X | Y$_0$}** to be approximately [1.0, 12.0].

Part B)ix) Analytical Calculation of Calibration CI/PI Intervals: T-cell Response = .3

$\hat{x} = \frac{Y_0 - \hat{\beta}_0}{\hat{\beta}_1}$   $\sqrt{MSE}\ (\sigma) = .08102$   SE(μ{Y|X = 6.47471}) = .0193   $T_{CRIT}$: $t_{.975,\ .025,\ 19} = \pm\ 2.093024$

**For $Y_0 = .3$** → $\frac{.3 - .00875}{.03282} = 6.47471$   $T_{STAT}$: $\hat{x} \pm t_{19} * SE_{CI}(\hat{x})$ → $6.47471 \pm 2.093024(.588056)$

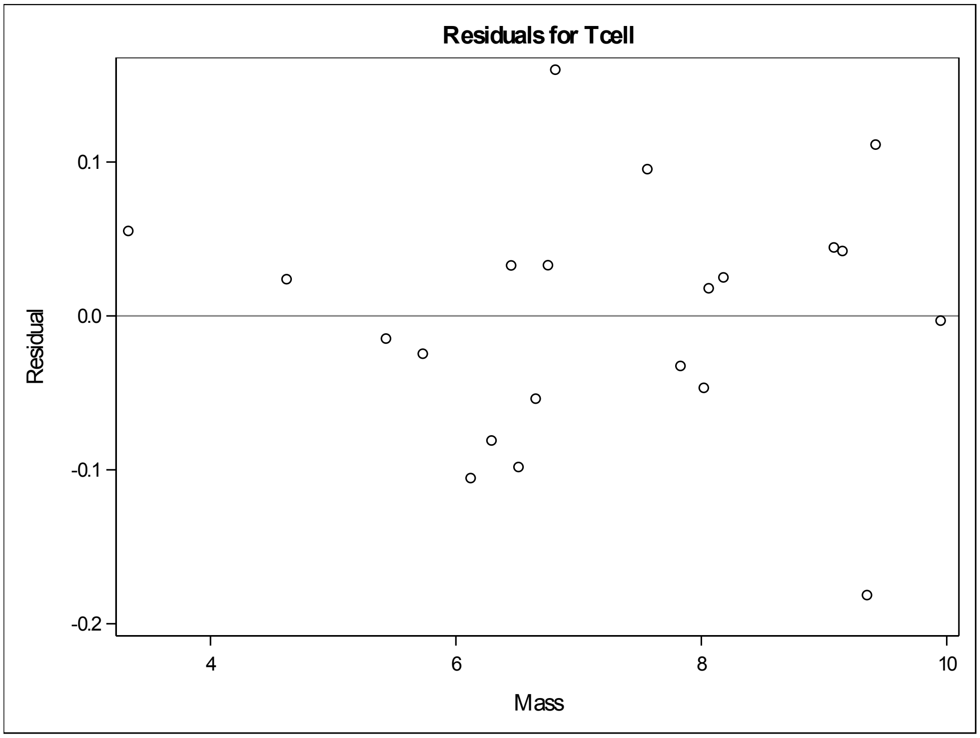
$SE_{CI}(\hat{x}) = \frac{SE(\mu\{Y|\hat{x}\})}{|\hat{\beta}_1|}$ → $\frac{.0193}{.03282} = .588056$   Inverse 95% CI {X|Y$_0$} → [5.243895, 7.705525]

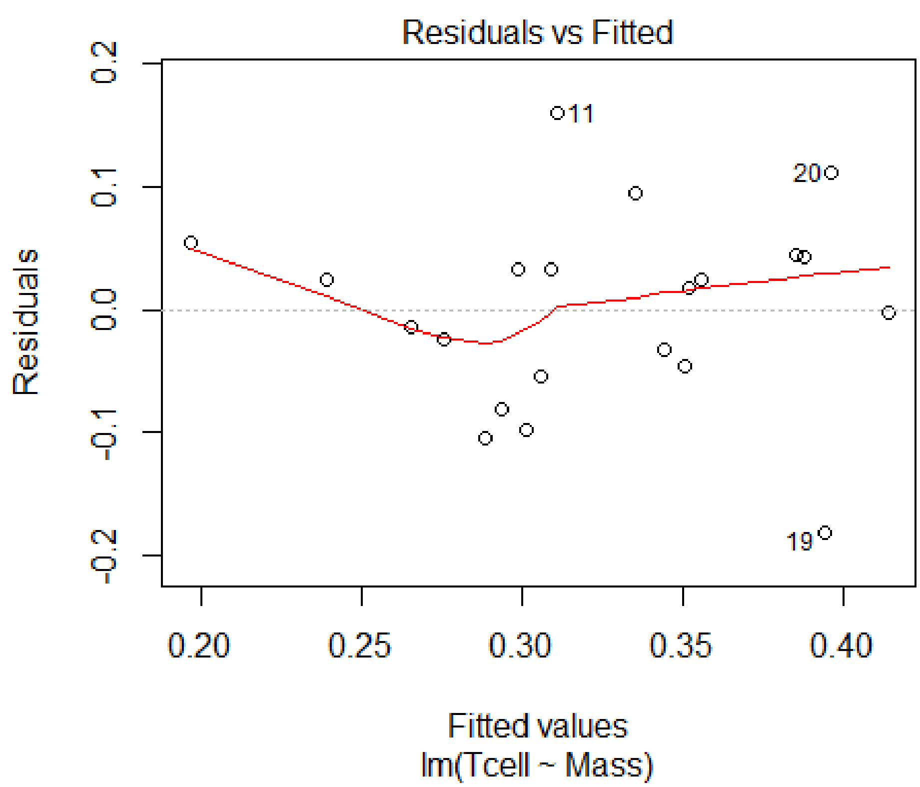
$SE_{PI}(\hat{x}) = \frac{SE(Pred\{Y|\hat{x}\})}{|\hat{\beta}_1|}$   SE(Pred{Y|X = 6.47471}) = $\sqrt{(.08102)^2 + (.0193)^2} = .0832868$

→ $\frac{.0832868}{.03282} = 2.537684$   $T_{STAT}$: $\hat{x} \pm t_{19} * SE_{PI}(\hat{x})$ → $6.47471 \pm 2.093024(2.537684)$

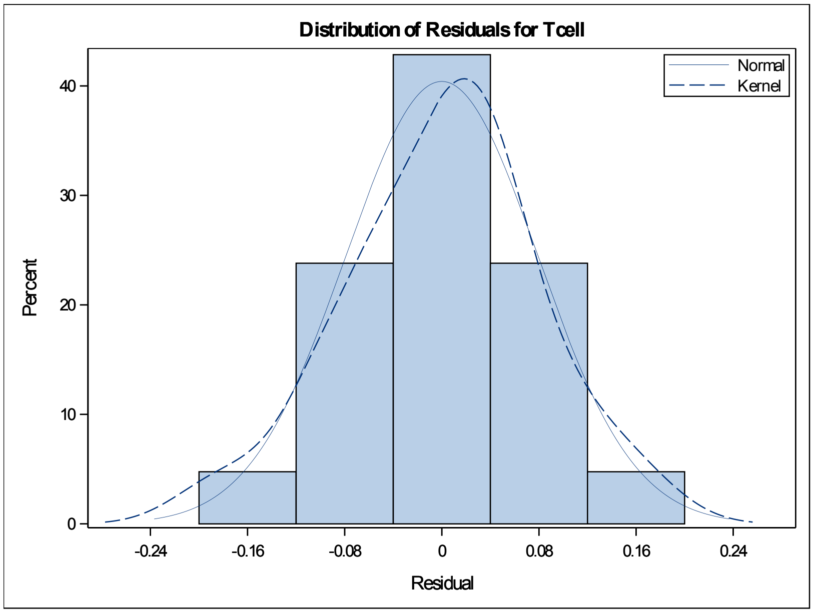
Inverse 95% PI {X|Y$_0$} → [1.163276, 11.786144]

Part B)x) Residual Scatterplots in SAS/R



Part B)xi) Residual Histogram with Superimposed Normal Distribution in SAS




**Question 2)** Parts A & B → **See Separate Excel Sheet Attachment!**

**Bonus**: Part C) **99% CI** for {Y|X=3} → [.04833, .3236] (Fitted Value: .186)

{Y|X=4} → [.1089, .3287] (Fitted Value: .2188)

{Y|X=5} → [.1676, .3356] (Fitted Value: .2516)

{Y|X=6} → [.222, .3469] (Fitted Value: .2844)

{Y|X=7} → [.2663, .3682] (Fitted Value: .3172)

{Y|X=8} → [.294, .4062] (Fitted Value: .3501)

{Y|X=9} → [.3084, .4574] (Fitted Value: .3829)

**Bonus**: Part D) **99% PI** for {Y|X=3} → [-.08361, .4555] (Fitted Value: .186)

{Y|X=4} → [-.03774, .4753] (Fitted Value: .2188)

{Y|X=5} → [.00505, .4982] (Fitted Value: .2516)

{Y|X=6} → [.04436, .5245] (Fitted Value: .2844)

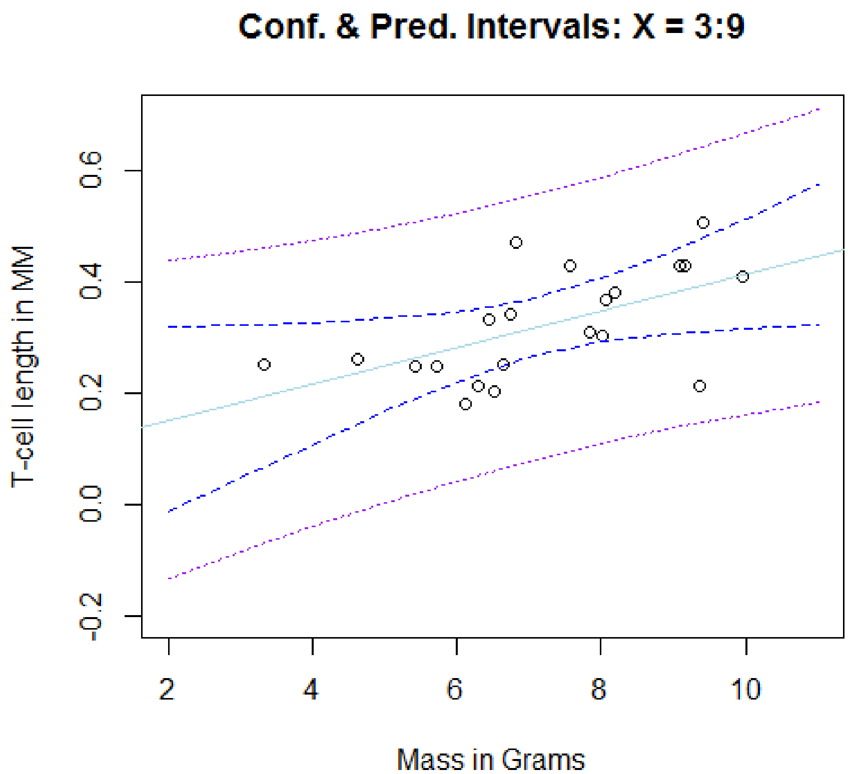
{Y|X=7} → [.07992, .5546] (Fitted Value: .3172)

{Y|X=8} → [.1116, .5885] (Fitted Value: .3501)

{Y|X=9} → [.1394, .6264] (Fitted Value: .3829)

Part E) →See Question 1, Parts VIII and IX → (**Repeated Question**)!

Part F) Plots for 99% Confidence & Prediction Intervals for Values between X=3:9

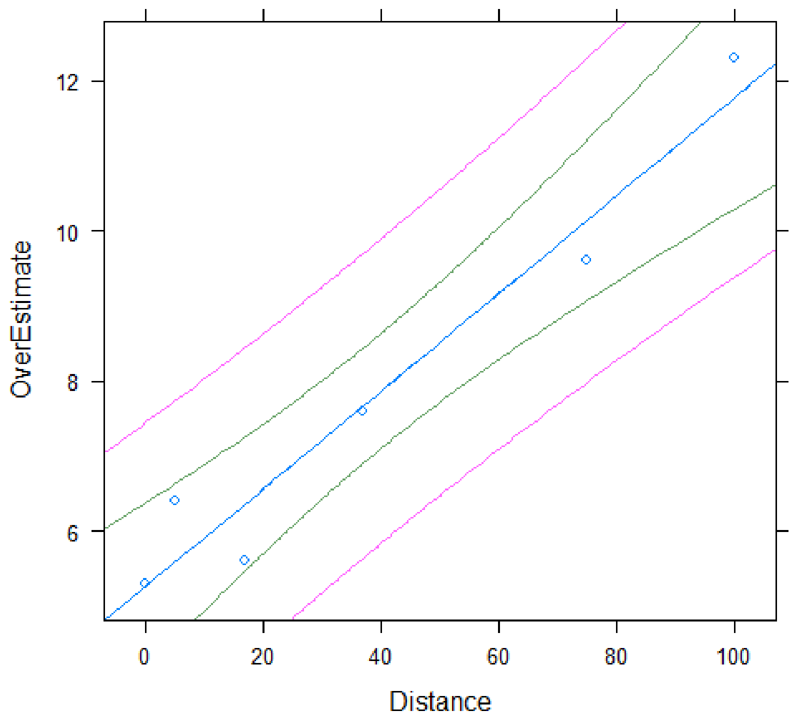


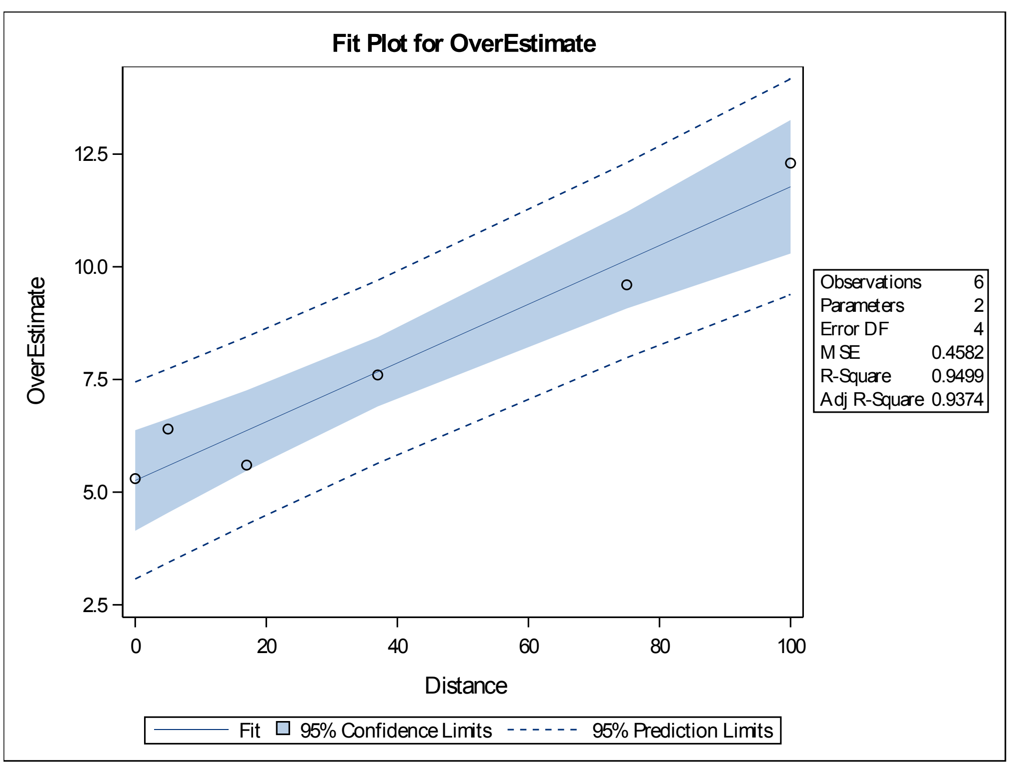

*R Code*:

```
> newer_X <- seq(2, 11, by = 1)

> plot(ex0727$Mass, ex0727$Tcell, ylim = c(-.2, .7), xlim = c(2, 11), xlab =
"Mass in Grams", ylab = "T-cell length in MM", main = "Conf. & Pred. Interval
s: X = 3:9")

> abline(lm_birds, col = "lightblue")

> conf_intv <- predict(lm_birds, newdata = data.frame(Mass = newer_X), interv
al = "confidence", level = .99)

> lines(newer_X, conf_intv[,2], col = "blue", lty = 2)

> lines(newer_X, conf_intv[,3], col = "blue", lty = 2)

> pred_intv <- predict(lm_birds, newdata = data.frame(Mass = newer_X), interv
al = "prediction", level = .99)

> lines(newer_X, pred_intv[,2], col = "purple", lty = 3)

> lines(newer_X, pred_intv[,3], col = "purple", lty = 3)
```

**Question 3)**

Part A)1) Scatterplots in R/SAS w/ Regression Confidence & Prediction Intervals



Part A)2) Tabular Regression Parameter T-Stats and P-values in SAS/R

| Parameter Estimates | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | 95% Confidence Limits | |
| Intercept | 1 | 5.25847 | 0.40192 | 13.08 | 0.0002 | 4.14257 | 6.37437 |
| Distance | 1 | 0.06517 | 0.00748 | 8.71 | 0.0010 | 0.04439 | 0.08594 |

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)

(Intercept)  5.25847    0.40192   13.08  0.00020 ***
Distance     0.06517    0.00748    8.71  0.00096 ***

---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Part A)3) Six-Step Hypothesis Tests for Regression Parameters

*Slope:*

$H_0$: $\beta_1 = 0$ $T_{CRIT} = t_{.975,\ .025,\ 4} = \pm\ 2.776445$

$H_A$: $\beta_1 \neq 0$ $T_{STAT} = \frac{.06517 - 0}{.00748} = 8.7125668$

$P_{VAL}$ = .000956 → Reject $H_0$

→ There is **sufficient** evidence at the $\propto$ = .05 level of significance ($P_{VAL}$ = .000956) to suggest that: the value of the slope parameter coefficient is not equal to zero.

*Intercept*:

$H_0$: $\beta_0 = 0$ $T_{CRIT} = t_{.975,\ .025,\ 4} = \pm\ 2.776445$

$H_A$: $\beta_0 \neq 0$ $T_{STAT} = \frac{5.25847 - 0}{.40192} = 13.083375$

$P_{VAL}$ = .000197 → Reject $H_0$

→ There is **sufficient** evidence at the $\propto$ = .05 level of significance ($P_{VAL}$ = .000197) to suggest that: the value of the intercept parameter coefficient is not equal to zero.

Part A)4) Regression Equation

*Model Estimate*: $\hat{y} = 5.25847 + .06517x$

Part A)5) Interpretation of Regression Parameters

*Parameter Interpretation*:

Slope: For every unit increase in Distance (ft.), the predicted OverEstimate (score) increases by .06517.

Intercept: As there was, in fact, a zero-valued observation of Distance (min: 0), the intercept is of some practical significance (indicating no distance between the interviewer and the door)—though it also signifies, from the quantitative perspective of the regression estimate, the point at which Distance no longer factors into the predicted OverEstimate value.

Part A)6) & 7) 95% CI/PI & Interpretation: OverEstimate Response at Distance = 37 ft. in R/SAS

*95% CI*: OverEstimate | Distance (37ft.) → [6.9013, 8.4380] → (Fitted Value = 7.6697)

95% of the repetitions of the sampling process include a **mean** response associated with the above intervals at the specified explanatory value of Distance = 37ft.

*95% PI*: OverEstimate | Distance (37ft.) → [5.6393, 9.7000] → (Fitted Value = 7.6697)

95% of the repetitions of the sampling process include an **individual** response associated with the above intervals at the specified explanatory value of Distance = 37ft.

```
> predict(lm_pol, interval = "confidence", data.frame(Distance = 37))

   fit   lwr   upr
1 7.67 6.901 8.438

> predict(lm_pol, interval = "prediction", data.frame(Distance = 37))

   fit   lwr upr
1 7.67 5.639 9.7
```

**Output Statistics**

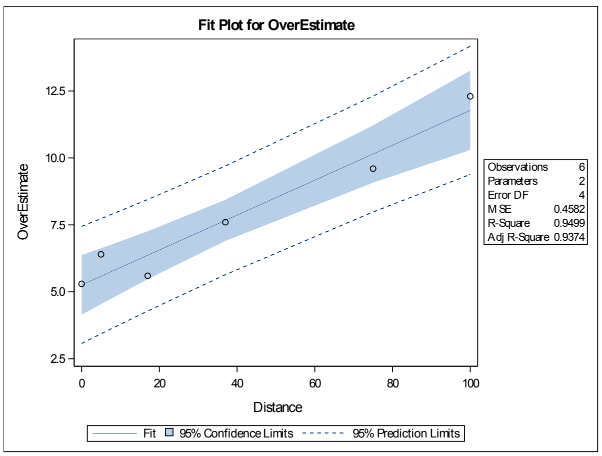
| Obs | Dependent Variable | Predicted Value | Std Error Mean Predict | 95% CL Mean | | 95% CL Predict | | Residual | Std Error Residual | Student Residual | Cook's D |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | . | 7.6697 | 0.2767 | 6.9013 | 8.4380 | 5.6393 | 9.7000 | . | . | . | . |

Part A)8) Software/Graphical Method: Calibration CI/PI Intervals: OverEstimate Response = 6 in R/SAS

```
> calibrate(lm_pol, 6, interval = "Wald", mean.response = TRUE)

estimate     lower     upper       se
  11.379    -3.324    26.081    5.295

> calibrate(lm_pol, 6, interval = "Wald", mean.response = FALSE)

estimate     lower     upper       se
   11.38    -20.99     43.75    11.66
```

*Interpretation*:

At the specified mean/individual OverEstimate value, the respective 95% CI/PI for the corresponding measurement of Distance would be [-3.324, 26.081] and [-20.99, 43.75], respectively.




Extending a horizontal line outward from **Y = 6** and two vertical lines down from its points of intersection with the indicated CI and PI bands, one can estimate the **inverse 95% CI {X | $Y_0$}** to be approximately [-1.0, 23.0] and the **inverse 95% PI {X | $Y_0$}** to be approximately [-20.0, 45.0].

Part A)9) Analytical Calculation of Calibration CI/PI Intervals: OverEstimate Response = 6

$\hat{x} = \frac{Y_0 - \hat{\beta}_0}{\hat{\beta}_1}$ $\sqrt{MSE}\ (\sigma) = .67689$ SE(μ{Y|X = 11.378395}) = .3451 $T_{CRIT}$: $t_{.975, .025, 4} = \pm 2.776445$

**For $Y_0 = 6$** → $\frac{6 - 5.25847}{.06517} = 11.378395$ $T_{STAT}$: $\hat{x} \pm t_4 * SE_{CI}(\hat{x})$ → $11.378395 \pm 2.776445(5.295381)$

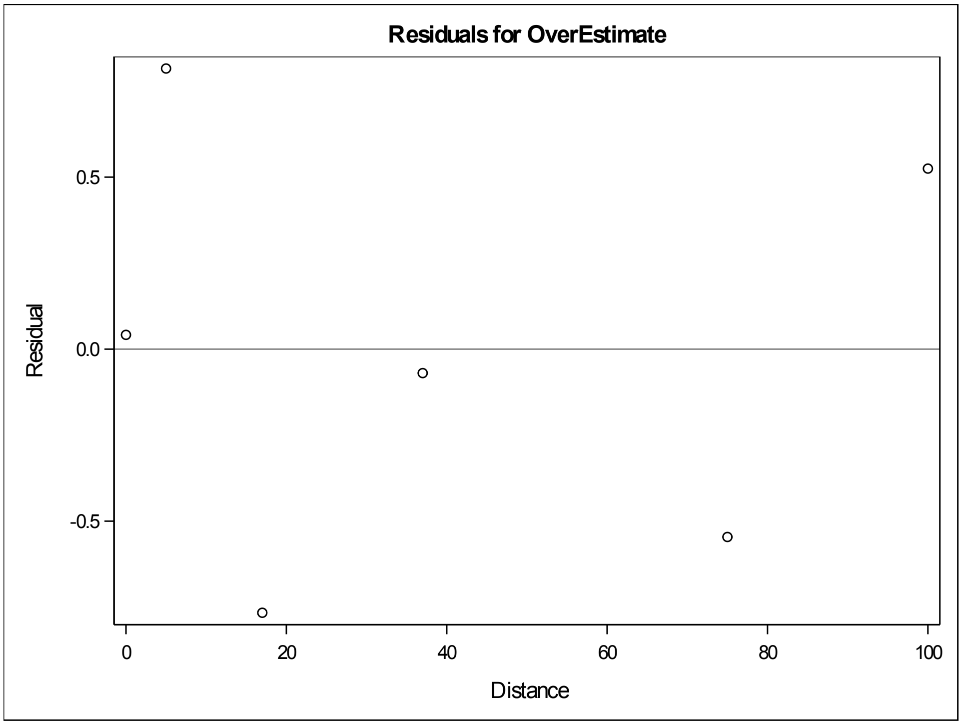
$SE_{CI}(\hat{x}) = \frac{SE(\mu\{Y|\hat{x}\})}{|\hat{\beta}_1|}$ → $\frac{.3451}{.06517} = 5.295381$ Inverse 95% CI {X|$Y_0$} → [-3.323939, 26.080729]

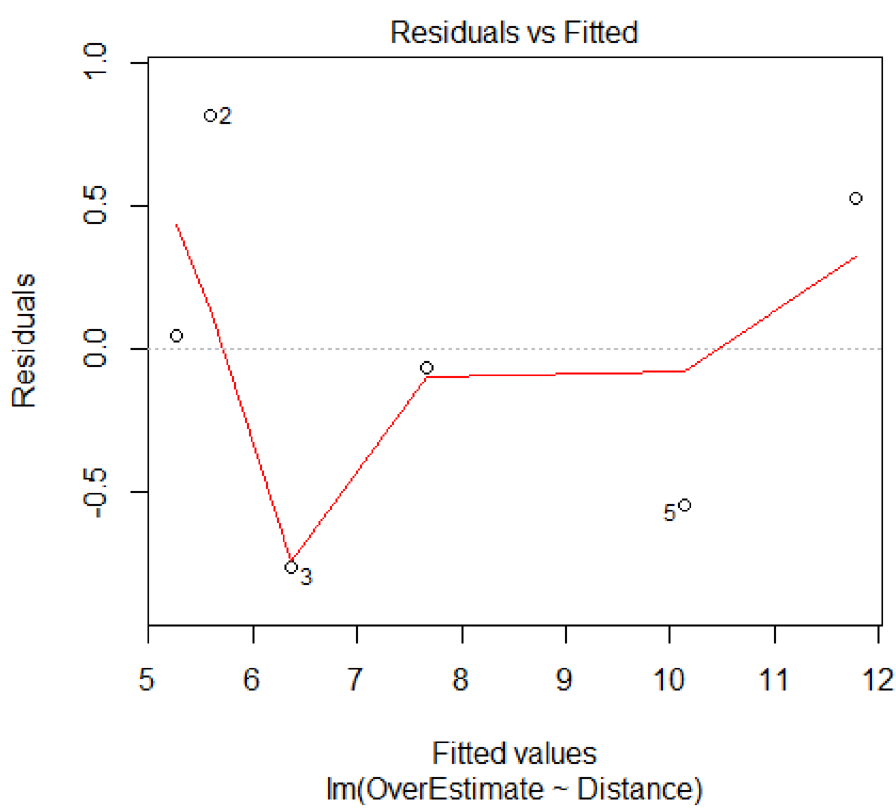
$SE_{PI}(\hat{x}) = \frac{SE(Pred\{Y|\hat{x}\})}{|\hat{\beta}_1|}$ SE(Pred{Y|X = 11.378395}) = $\sqrt{(.67689)^2 + (.3451)^2} = .7597856$

→ $\frac{.7597856}{.06517} = 11.658517$ $T_{STAT}$: $\hat{x} \pm t_4 * SE_{PI}(\hat{x})$ → $11.378395 \pm 2.776445(11.658517)$

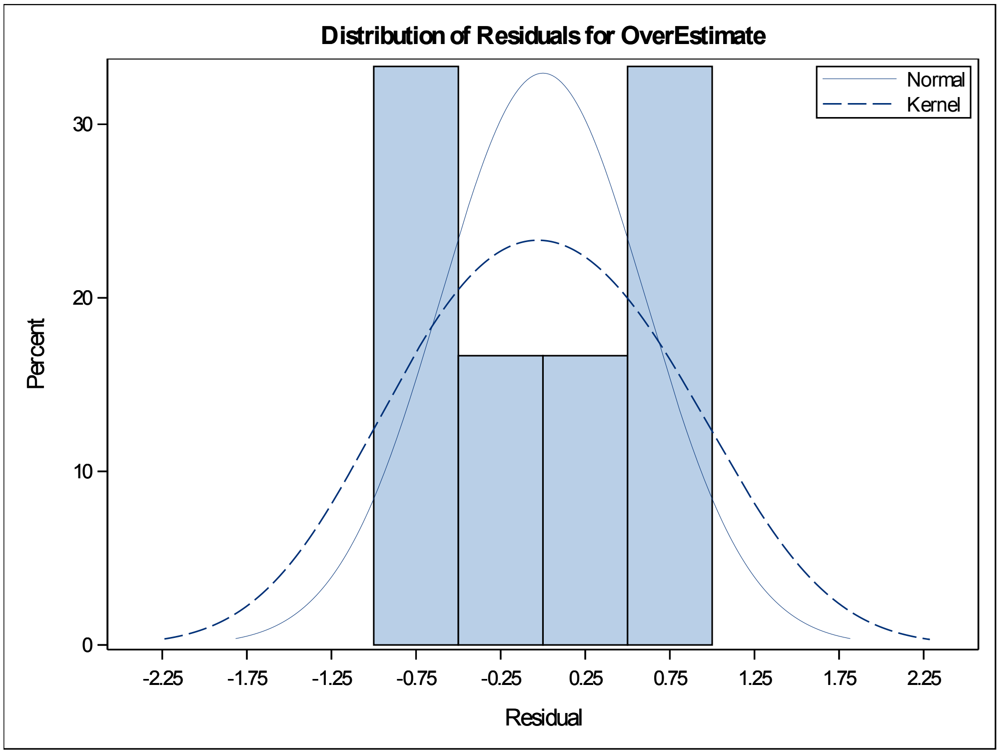
Inverse 95% PI {X|$Y_0$} → [-20.99084, 43.7476262]

Part A)10) Residual Scatterplots in SAS/R



Part A)11) Residual Histogram with Superimposed Normal Distribution in SAS




Part B) Answer to Textual Question [Stat Sleuth pp. 205-206]

On the basis of the significant parameter coefficient estimates, the graphical output, and the value of $R^2$, one can (given the *small sample size*) conservatively say there is substantial statistical evidence that the mean OverEstimate increases with increasing Distance of the interviewer from the door.

Bonus Question)

The unit of measurement for Distance used in the study was **feet (ft.)**.

**Question 4)**

For Question 1) Measure of Variation in the Response Accounted for by the Explanatory Variable

$R^2 = 100(\frac{SST - SSR}{SST})\% = \frac{.18716 - .12472}{.18716} = 33.36183\%$

→ Approximately 33.36% of the variation in the response (T-cell telomere length in mm.) was explained by the linear regression on the explanatory variable (mean stone Mass in grams).

*Tabular Verification of $R^2$ in SAS/R*:

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Model** | 1 | 0.06244401 | 0.06244401 | 9.51 | 0.0061 |
| **Error** | 19 | 0.12471894 | 0.00656415 | | |
| **Corrected Total** | 20 | 0.18716295 | | | |

| R-Square | Coeff Var | Root MSE | Tcell Mean |
|---|---|---|---|
| 0.333634 | 25.00969 | 0.081019 | 0.323952 |

```
> summary(lm_birds)

Call:
lm(formula = Tcell ~ Mass, data = ex0727)

…

Residual standard error: 0.081 on 19 degrees of freedom
Multiple R-squared:  0.334,    Adjusted R-squared:  0.299
F-statistic: 9.51 on 1 and 19 DF,  p-value: 0.00611

> SSR = sum(resid(lm_birds)^2)
> SSR
[1] 0.1247

> SSE = sum((fitted(lm_birds) - mean(~Tcell, data = ex0727))^2)
> SSE
[1] 0.06244

> 1 - (SSR/(SSE + SSR))
[1] 0.3336
```

<u>For Question 3) Measure of Variation in the Response Accounted for by the Explanatory Variable</u>

$R^2 = 100(\frac{SST - SSR}{SST})\% = \frac{36.58-1.83272061}{36.58} = 94.9898\%$

→ Approximately 94.99% of the variation in the response (OverEstimate point score) was explained by the linear regression on the explanatory variable (Distance of the interviewer from the door in ft.).

*Tabular Verification of $R^2$ in SAS/R*:

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 1 | 34.74727939 | 34.74727939 | 75.84 | 0.0010 |
| Error | 4 | 1.83272061 | 0.45818015 | | |
| Corrected Total | 5 | 36.58000000 | | | |

| R-Square | Coeff Var | Root MSE | OverEstimate Mean |
|---|---|---|---|
| 0.949898 | 8.678078 | 0.676890 | 7.800000 |

```
> summary(lm_pol)

Call:
lm(formula = OverEstimate ~ Distance, data = ex0729)

…

Residual standard error: 0.677 on 4 degrees of freedom
Multiple R-squared:  0.95,	Adjusted R-squared:  0.937
F-statistic: 75.8 on 1 and 4 DF,  p-value: 0.000957


> SSR_pol = sum(resid(lm_pol)^2)
> SSR_pol
[1] 1.833

> SSE_pol = sum((fitted(lm_pol) - mean(~OverEstimate, data = ex0729))^2)
> SSE_pol
[1] 34.75

> 1 - (SSR_pol/(SSE_pol + SSR_pol))
[1] 0.9499
```

***R/SAS Code*** **:**

*R* -- Question 1)

```
> require(Sleuth3)
> require(mosaic)
> options(digits = 4)
> summary(ex0727)
> lm_birds = lm(Tcell ~ Mass, data = ex0727)
> summary(lm_birds)
> xyplot(Tcell ~ Mass, panel = panel.lmbands, data = ex0727)
> predict(lm_birds, interval = "confidence", data.frame(Mass = 4.5))
> predict(lm_birds, interval = "prediction", data.frame(Mass = 4.5))
> library(investr)
> calibrate(lm_birds, .3, interval = "Wald", mean.response = TRUE)
> calibrate(lm_pol, 6, interval = "Wald", mean.response = FALSE)
> plot(lm_birds, which = 1)
> SSR = sum(resid(lm_birds)^2)
> SSE = sum((fitted(lm_birds) - mean(~Tcell, data = ex0727))^2)
> 1 - (SSR/(SSE + SSR))
```

*R* -- Question 3)

```
> require(Sleuth3)
> require(mosaic)
> options(digits = 4)
> summary(ex0729)
> lm_pol = lm(OverEstimate ~ Distance, data = ex0729)
> summary(lm_pol)
> xyplot(OverEstimate ~ Distance, panel = panel.lmbands, data = ex0729)
> predict(lm_pol, interval = "confidence", data.frame(Distance = 37))
> predict(lm_pol, interval = "prediction", data.frame(Distance = 37))
> library(investr)
> calibrate(lm_pol, 6, interval = "Wald", mean.response = TRUE)
> calibrate(lm_pol, 6, interval = "Wald", mean.response = FALSE)
> plot(lm_pol, which = 1)
> SSR_pol = sum(resid(lm_pol)^2)
> SSE_pol = sum((fitted(lm_pol) - mean(~OverEstimate, data = ex0729))^2)
> 1 - (SSR_pol/(SSE_pol + SSR_pol))
```

<u>*SAS* -- Question 1)</u>

```
FILENAME REFFILE '/home/jrasmusvorrath0/ex0727.csv';

PROC IMPORT DATAFILE=REFFILE

        DBMS=CSV

        OUT=WORK.IMPORT3;

        GETNAMES=YES; RUN;

PROC CONTENTS DATA=WORK.IMPORT3; RUN;

data birds; set work.import3; run;

proc print data = birds; run;

proc corr data = birds plots = scatter;

        var Mass Tcell; run;

proc reg data = birds plots = residuals;

        model Tcell = Mass / R CLM CLI CLB; run;

        quit;

---

data birds_pred;

        input Mass Tcell;

        datalines;

4.5 na

3.33 .252

4.62 .263

5.43 .251

5.73 .251

6.12 .183

6.29 .213
```

```
6.45 .332

6.51 .203

6.65 .252

6.75 .342

6.81 .471

7.56 .431

7.83 .312

8.02 .304

8.06 .37

8.18 .381

9.08 .43

9.15 .43

9.35 .213

9.42 .508

9.95 .411; run;

proc print data = birds_pred; run;

proc glm data = birds_pred;

	model Tcell = Mass / clparm;

	output out = birds_pred_resid residual = Residuals; run;

*proc print data = birds_pred_resid; *run;

*proc means data = birds_pred_resid var;

	*var Tcell Residuals; *run;

proc reg data = birds_pred;

	model Tcell = Mass / ss1 ss2 clb stb r cli clm; run;

data birds_temp;
```

```
	input Mass;
	datalines;
6.47471; run;
data birds_pred_2;
	set birds_pred birds_temp; run;
proc print data= birds_pred_2; run;
proc reg data= birds_pred_2;
	model Tcell = Mass /CLB CLM CLI; run;
	quit;
```

*SAS* -- Question 3)

```
FILENAME REFFILE '/home/jrasmusvorrath0/ex0729.csv';
PROC IMPORT DATAFILE=REFFILE
	DBMS=CSV
	OUT=WORK.IMPORT2;
	GETNAMES=YES; RUN;
PROC CONTENTS DATA=WORK.IMPORT2; RUN;
data pol;
	set work.import2; run;
proc print data = pol; run;
proc corr data = pol plots = scatter;
	var Distance OverEstimate; run;
proc reg data = pol plots = residuals;
	model OverEstimate = Distance / R CLM CLI CLB; run;
	quit;
```

---

```
data pol_pred;
	input OverEstimate Distance;
	datalines;
na 37
5.3 0
6.4 5
5.6 17
7.6 37
9.6 75
12.3 100; run;
proc print data = pol_pred; run;
proc glm data = pol_pred;
	model OverEstimate = Distance / clparm;
	output out = pol_pred_resid residual = Residuals; run;
*proc print data = pol_pred_resid; *run;
*proc means data = pol_pred_resid var;
	*var OverEstimate Residuals; *run;
proc reg data = pol_pred;
	model OverEstimate = Distance / ss1 ss2 clb stb r cli clm; run;
data pol_temp;
	input Distance;
	datalines;
11.378395; run;
data pol_pred_2;
```

```
set pol_pred pol_temp; run;

proc print data= pol_pred_2; run;

proc reg data= pol_pred_2;

    model OverEstimate = Distance /CLB CLM CLI; run;

    quit;
```

**Bonus Question)**

Part A)

Setting partial derivatives of SS($\beta_0, \beta_1$) with respect to each parameter equal to zero, to show that $\beta_0$ and $\beta_1$ must satisfy the normal equations:

$$\beta_0 n + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$Y_i = b_0 + b_1 x_i + e_i$ SSE = $\sum_{i=1}^{n} e_i^2$

$e_i = Y_i - b_0 + b_1 x_i$

**SSE** = $\boldsymbol{\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2}$

$$\frac{\partial}{\partial b_0}\text{SS}(\beta_0, \beta_1) = \frac{\partial}{\partial b_0}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$$

$$0 = 2\left(\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)\right)$$

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}(b_0 + b_1 x_i)$$

$$= \sum_{i=1}^{n} b_0 + \sum_{i=1}^{n}(b_1 x_i)$$

$$= b_0 n + b_1 \sum_{i=1}^{n} x_i$$

$$\frac{\partial}{\partial b_1}\text{SS}(\beta_0, \beta_1) = \frac{\partial}{\partial b_1}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$$

$$0 = 2\left(\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)\right) x_i$$

$$= \sum_{i=1}^{n}\left(y_i x_i - b_0 x_i - b_1 x_i^2\right)$$

$$\sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n}(b_0 x_i + b_1 x_i^2)$$

$$= b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$

Part B)

$$\sum_{i=1}^{n} y_i = b_0 n + b_1 \sum_{i=1}^{n} x_i$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i = b_0 + b_1\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)$$

$$\bar{Y} = B_0 + B_1 \bar{X}$$

$$B_0 = \bar{Y} - B_1 \bar{X}$$

$$B_0 = \bar{Y} - B_1 \bar{X}$$

$$\sum_{i=1}^{n} Y_i X_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i x_i = (\bar{Y} - B_1 \bar{X})\frac{1}{n}\sum_{i=1}^{n} x_i + b_1 \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

$$B_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Part C)

$B_0 = \bar{Y} - B_1\bar{X}$ $B_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

$$\begin{aligned}
SS(\beta_0, \beta_1) &= \sum_{i=1}^{n}(y_i - b_0 - b_1x_i)^2 \\
&= \boldsymbol{\sum_{i=1}^{n}(y_i - \bar{y} + b_1\bar{x} - b_1x_i)^2} \\
&= \boldsymbol{\sum_{i=1}^{n}(y_i - \bar{y} - b_1(x_i - \bar{x}))^2} \\
&= \boldsymbol{\sum_{i=1}^{n}((y_i - \bar{y}) - \left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)(x_i - \bar{x}))^2} \\
&= 0
\end{aligned}$$

∴ The solutions give minimum values to the Sum of Squares (SS($\beta_0, \beta_1$)).

Furthermore, taking second derivatives:

$\frac{\partial^2}{\partial b_0^2}$ = **2n** $\frac{\partial^2}{\partial b_1^2}$ = **2n**$\boldsymbol{x_i^2}$

∴ Values of $\beta_0$, $\beta_1$ that satisfy equations generated by setting partial derivatives equal to zero refer to minimum values of SSE.