Formulas

$B_0 = \bar{y} - b_1\bar{x}$ $B_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$

$S = \sqrt{\frac{\sum(y_i - \hat{y})^2}{n-2}}$ $SSR = \sum(y_i - \hat{y})^2 = \sum(y_i - \bar{y})^2 - b_1^2\sum(x_i - \bar{x})^2$

$SE(b_0) = s\sqrt{\frac{\sum x_i^2}{n\sum(x_i - \bar{x})^2}}$ $SE(b_1) = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}$

Question 1: Summary Statistics

$\sum_{i=1}^{30} x_i = 2707$ $\sum_{i=1}^{30} x_i^2 = 286509$ $\sum_{i=1}^{30} x_i y_i = 223728$ $\sum_{i=1}^{30}(x_i - \bar{x})^2 = 42247.37$

$\sum_{i=1}^{30} y_i = 2430$ $\sum_{i=1}^{30} y_i^2 = 200342$ $\sum_{i=1}^{30}(y_i - \bar{y})^2 = 3512$ $\sum_{i=1}^{30}(x_i - \bar{x})(y_i - \bar{y}) = 4461$

Question 1)

Part a) Finding the Least Squares Regression line for predicting Wins | Payroll

Calculations:

$B_1 = \frac{4461}{42247.37} = .105592$ $B_0 = \frac{2430}{30} - .105592(\frac{2707}{30}) = 71.472082$

$SSR = 3512 - (.105592)^2 \times (42247.37) = 3040.95575$ $S = \sqrt{\frac{3040.95575}{28}} = 10.421399$

$SE(b_1) = \frac{10.421399}{\sqrt{42247.37}} = .050702$ $SE(b_0) = 10.421399\sqrt{\frac{286509}{30(42247.37)}} = 4.954898$

Model Estimate: $\hat{y} = 71.472082 + .105592x$

Parameter Interpretation: For every extra $1M in Payroll, the predicted # of wins increases by .105592

Part b) Six-Step Hypothesis Test for Slope Parameter

H_0: $\beta_1 = 0$ $T_{CRIT} = t_{.975, .025, 28} = \pm 2.048407$

H_A: $\beta_1 \neq 0$ $T_{STAT} = t_{28} = \frac{.1056 - 0}{.050702} = 2.082758$

$P_{VAL} = .046525$ → Reject H_0

→ There is sufficient evidence at the $\propto$ = .05 level of significance (P_{VAL} = .046525) to suggest that: the value of the slope parameter coefficient is not equal to zero.

Part c) Confidence Interval for Slope Parameter

95% CI (slope): $b_1 \pm t_{28} * SE(b_1) = .105592 \pm 2.048407(.050702) \rightarrow$ [.001734 , .20945]

→The calculated parameter confidence interval, which does not include zero, is consistent with the result (Reject H_0) of the hypothesis test for H_0: $\beta_1 = 0$.

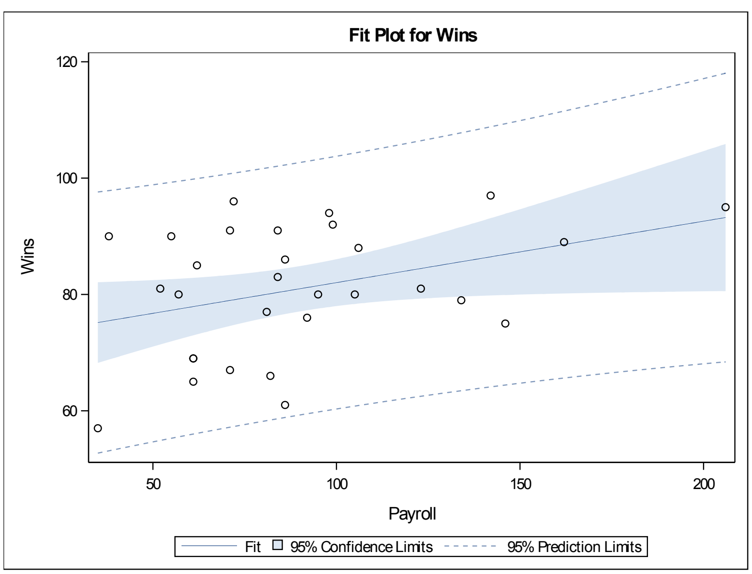
Part d) SAS/R Results and Code

SAS Results:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	471.047608	471.047608	4.34	0.0465
Error	28	3040.952392	108.605443		
Corrected Total	29	3512.000000			

R-Square	Coeff Var	Root MSE	Wins Mean
0.134125	12.86592	10.42139	81.00000

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	71.47204757	4.95489528	14.42	<.0001	61.32240470	81.62169044
Payroll	0.10559238	0.05070210	2.08	0.0465	0.00173383	0.20945093

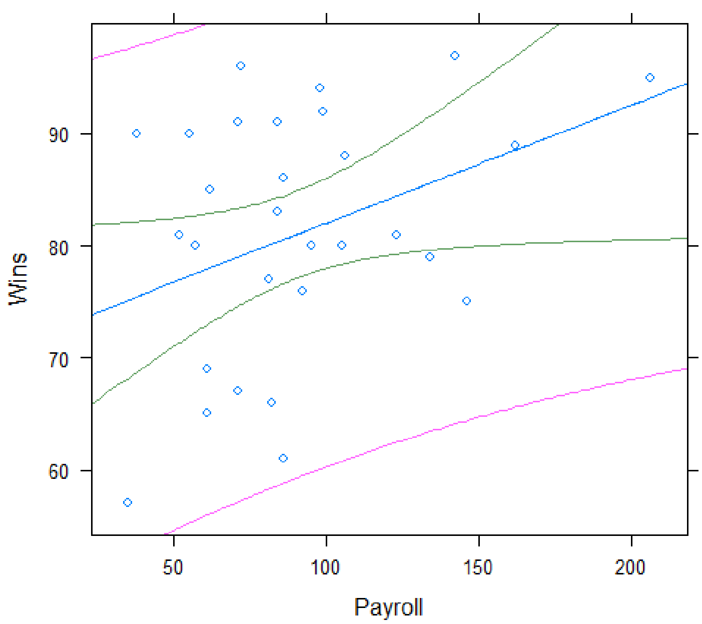


SAS Code:

```
FILENAME REFFILE '/home/jrasmusvorrath0/baseball - Payroll_Wins_2010.xlsx';

PROC IMPORT DATAFILE=REFFILE

        DBMS=XLSX

        OUT=WORK.IMPORT2;

        GETNAMES=YES; RUN;


PROC CONTENTS DATA=WORK.IMPORT2; RUN;

data baseballz; set work.import2; run;

proc print data = baseballz; run;


proc glm data = baseballz plots(unpack)= diagnostics;

model wins = payroll / clparm;

output out = baseballz_resid residual = Residuals; run;


*proc print data = baseballz_resid; run;


*proc means data = baseballz_resid var;

*var wins Residuals; run;


proc reg data= baseballz;

model wins = payroll / ss1 ss2 clb stb r cli clm; run;
```

R Results:

```
Call:
lm(formula = Wins ~ Payroll, data = baseball4)

Residuals:
   Min     1Q Median     3Q    Max 
-19.55  -8.34   1.10   9.30  16.93 

Coefficients:
            Estimate Std. Error t value
(Intercept)  71.4720     4.9549   14.42
Payroll       0.1056     0.0507    2.08
            Pr(>|t|)    
(Intercept)  1.7e-14 ***
Payroll        0.047 *  
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.4 on 28 degrees of freedom
Multiple R-squared:  0.134,	Adjusted R-squared:  0.103 
F-statistic: 4.34 on 1 and 28 DF,  p-value: 0.0465

> sum(resid(lm4)^2)
[1] 3041

> confint(lm4)
                2.5 %  97.5 %
(Intercept) 61.322405 81.6217
Payroll      0.001734  0.2095
```



R Code:

```
> attach(baseball4)

> require(mosaic)

> options(digits = 4)

> xyplot(Wins ~ Payroll, type = c("p", "r"), data = baseball4)

> lm4 = lm(Wins ~ Payroll, data = baseball4)

> summary(lm4)

> resid(lm4)^2

> sum(resid(lm4)^2)

> confint(lm4)

> xyplot(Wins ~ Payroll, panel = panel.lmbands, data = baseball4)
```

Question 2)

Part a) Finding the Least Squares Regression line for predicting Math | Science

Model Estimate: $\hat{y} = 21.700192 + .596814x$

Parameter Interpretation:

Slope: For every extra point scored on Science, the predicted score of Math increases by .596814.

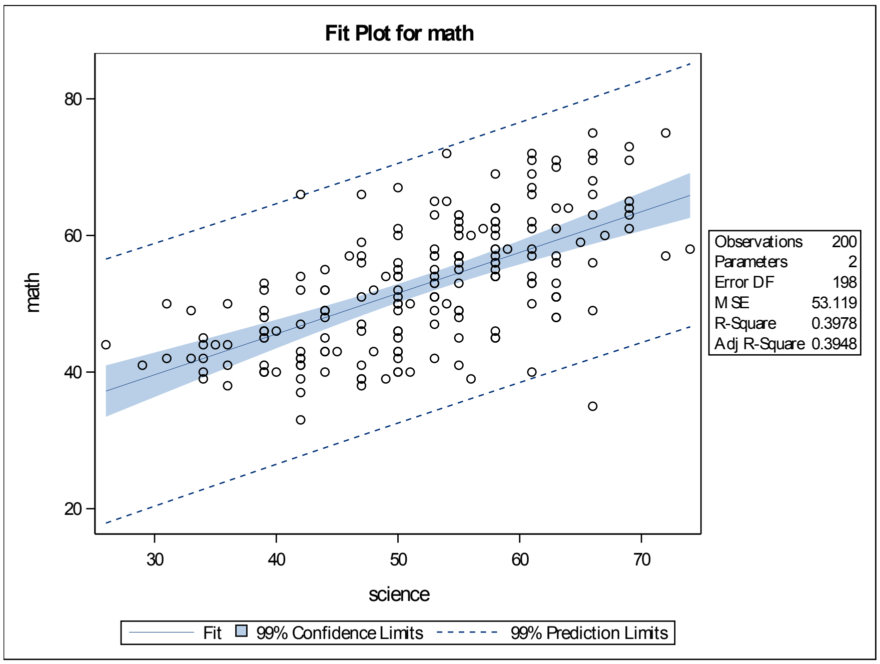
Intercept: As there were no zero-valued Science scores (min: 26), the intercept is not of practical significance, though it could be interpreted, from the quantitative perspective of the regression estimate, as the point at which the Science score no longer factors into the predicted Math score.

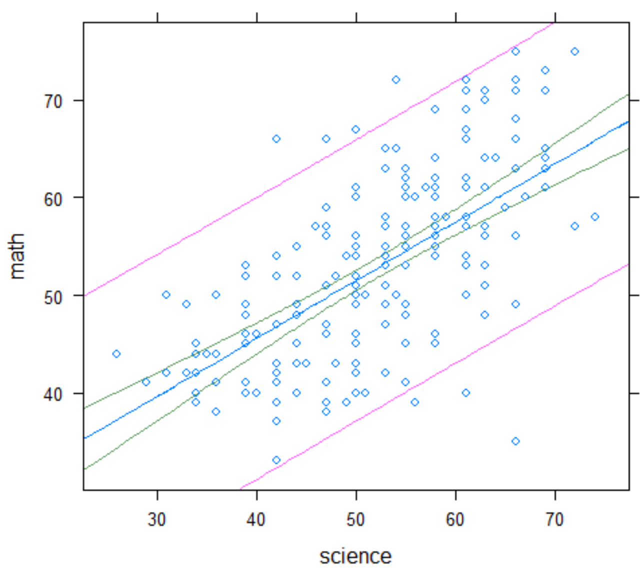
Results:

Parameter	Estimate	Standard Error	t Value	Pr > |t|	99% Confidence Limits	
Intercept	21.70019172	2.75429099	7.88	<.0001	14.53659134	28.86379211
science	0.59681405	0.05218220	11.44	<.0001	0.46109403	0.73253407

```
Call:
lm(formula = math ~ science, data = `hsb2.(1)`)

Residuals:
    Min      1Q  Median      3Q     Max
-26.090  -5.004   0.467   4.689  19.234
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.7002     2.7543    7.88  2.2e-13 ***
science       0.5968     0.0522   11.44  < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.29 on 198 degrees of freedom
Multiple R-squared:  0.398,   Adjusted R-squared:  0.395
F-statistic:  131 on 1 and 198 DF,  p-value: <2e-16
```





SAS Code:

```
FILENAME REFFILE '/home/jrasmusvorrath0/hsb2 (1).csv';

PROC IMPORT DATAFILE=REFFILE

        DBMS=CSV

        OUT=WORK.IMPORT1;

        GETNAMES=YES; RUN;

PROC CONTENTS DATA=WORK.IMPORT1; RUN;

data math_sci; set work.import1; run;

proc print data = math_sci; run;


*proc means data = math_sci;

*var science; run;


proc glm data = math_sci alpha = .01 plots(unpack)= diagnostics;

model math = science / clparm;

output out = math_sci_res residual = Residuals;

run;


*proc print data = math_sci_res; run;


*proc means data = math_sci_res var;

*var math Residuals; run;


proc reg data= math_sci alpha = .01;

model math = science / ss1 ss2 clb stb; run;
```

R Code:

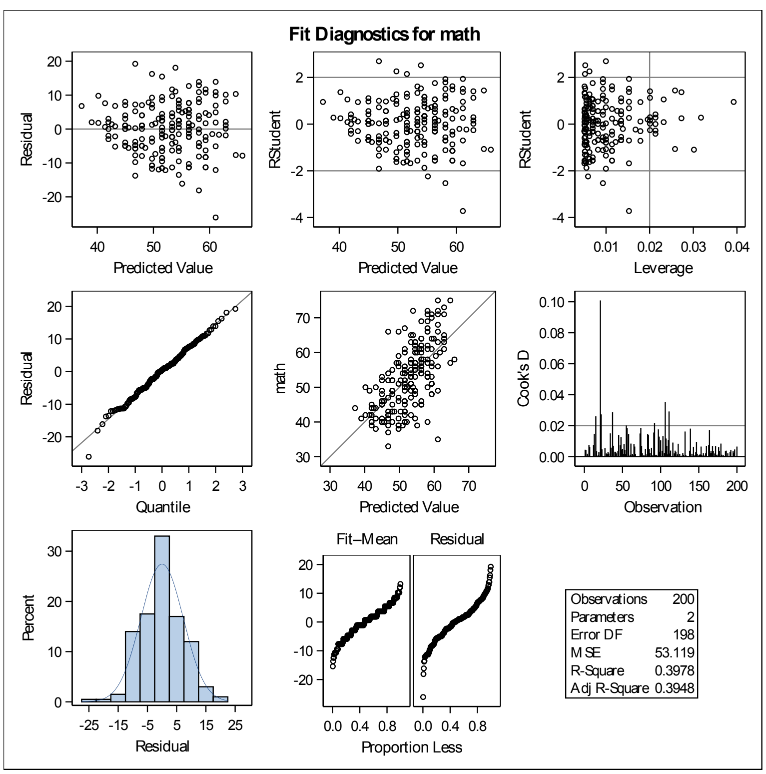
```
> `hsb2.(1)` <- read.csv("C:/Users/Jack/Desktop/M.S. Application Documents/SM
U/Courses/Experimental Statistics I/Data Sets/hsb2 (1).csv")

> View(`hsb2.(1)`)

> lm5 = lm(math ~ science, data = `hsb2.(1)`)

> summary(lm5)

> resid(lm5)^2

> sum(resid(lm5)^2)

> confint(lm5, level = .99)

> xyplot(math ~ science, panel = panel.lmbands, data = `hsb2.(1)`)
```

Part b) Six-Step Hypothesis Test for Slope (Science) & Intercept Parameters

Regression Assumptions:

On the basis of the study description, one can assume independence of errors. With the exception of a couple moderate outliers, a closer look at the residual QQ plots and histograms (Figure 1) confirms a normal distribution of these errors, with reasonably constant variance and linearity.

Figure 1:



Slope:

H_0: $\beta_1 = 0$ $T_{CRIT} = t_{.995, .005, 198} = \pm 2.600887$

H_A: $\beta_1 \neq 0$ $T_{STAT} = t_{198} = \frac{.596814 - 0}{.0521822} = 11.437118$

$P_{VAL} = .0001$ → Reject H_0

→ There is sufficient evidence at the $\propto = .01$ level of significance ($P_{VAL} = .0001$) to suggest that: the value of the slope parameter coefficient is not equal to zero.

Intercept:

H_0: $\beta_0 = 0$ $T_{CRIT} = t_{.995, .005, 198} = \pm 2.600887$

H_A: $\beta_0 \neq 0$ $T_{STAT} = t_{198} = \frac{21.700192 - 0}{2.754291} = 7.878685$

$P_{VAL} = .0001$ → Reject H_0

→ There is sufficient evidence at the $\propto = .01$ level of significance ($P_{VAL} = .0001$) to suggest that: the value of the intercept parameter coefficient is not equal to zero.

SAS Procedures:

```
proc glm data = math_sci alpha = .01 plots(unpack)= diagnostics;

model math = science / clparm;

output out = math_sci_res residual = Residuals; run;

proc reg data= math_sci alpha = .01;

model math = science / ss1 ss2 clb stb; run;
```

R Commands:

```
> lm5 = lm(math ~ science, data = `hsb2.(1)`)

> summary(lm5)

> xyplot(math ~ science, panel = panel.lmbands, data = `hsb2.(1)`)
```

Part c) Confidence Interval for Slope Parameter

99% CI (slope): $b_1 \pm t_{198} * SE(b_1) = .596814 \pm 2.600887\,(.0521822)$ → [.461094, .732534]

→The calculated parameter confidence interval, which does not include zero, is consistent with the result (Reject H_0) of the hypothesis test for H_0: $\beta_1 = 0$.

99% CI (intercept): $b_0 \pm t_{198} * SE(b_o) = 21.700192 \pm 2.600887\,(2.754291)$ → [14.536592, 28.863792]

→The calculated parameter confidence interval, which does not include zero, is consistent with the result (Reject H_0) of the hypothesis test for H_0: $\beta_0 = 0$.

Part d) SAS/R Confidence Interval Results and Code

SAS:

Parameter Estimates										
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS	Standardized Estimate	99% Confidence Limits	
Intercept	1	21.70019	2.75429	7.88	<.0001	554299	3297.26521	0	14.53659	28.86379
science	1	0.59681	0.05218	11.44	<.0001	6948.31801	6948.31801	0.63073	0.46109	0.73253

```
proc reg data= math_sci alpha = .01;

model math = science / ss1 ss2 clb stb; run;
```

R:

```
              0.5 %  99.5 %
(Intercept) 14.5366 28.8638
science      0.4611  0.7325

> lm5 = lm(math ~ science, data = `hsb2.(1)`)

> confint(lm5, level = .99)
```

Bonus Question) 95% Confidence & Prediction Interval for Wins | Payroll ($100M)

95% CI: Wins | Payroll ($100M) → [78.0040, 86.0586] → (Fitted Value = 82.0313)

95% PI: Wins | Payroll ($100M) → [60.3075, 103.7551] → (Fitted Value = 82.0313)

→ The 95% CI is constructed such that 95% of the repetitions of the sampling process result in intervals that include the correct mean response at a specified value of X.

→ Contrastingly, the wider 95% PI—which indicates likely values for a future value of a response variable at a specified value of X—is a measure of the likelihood that the interval will include the future response value. The PI not only factors in uncertainty about a parameter measurement (i.e., the subpopulation mean), but also uncertainty about an individual future value in relation to its mean.