ORIGIN ≡ 0

Inferences on "Simple" Linear Regression

W. Stein

Regression inferences are based on Simple Linear Regression involve devising statistical tests, or setting confidence intervals, on regression parameters (β_0 or β_1), expected mean response (Y_h), or predictions on new observations (Y_{hnew}). One may also set confidence bands for multiple simultaneous Y_h given fixed independent X_h, or on the entire regression. Terminology and Page Numbers drawn from Kuter et al. (KNNL) A*pplied Linear Statistical Models* 5th Edition.

Assumptions:

- Standard Linear Regression depends on specifying in advance which variable is to be considered 'dependent' and which 'independent'. This decision matters as changing roles for Y & X usually produces a different result.

- Y_1, Y_2, Y_3, ... , Y_n (dependent variable) is a random sample ~ $N(\mu,\sigma^2)$.

- X_1, X_2, X_3, ... , X_n (independent variable) with each value of X_i matched to Y_i

Model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where: β_0 is the y intercept of the regression line (translation),

β_1 is the slope of the regression line (scaling coefficient),

ε_i is the error factor in prediction of Y_i and a random variable distributed as $N(0,\sigma^2)$.

Note that although the Normality assumption is not strictly necessary in estimation of regression coefficients (β_i) using Least Squares Estimation, it applies in most other situations such as when using Maximum Liklihood Estimation, or statistical inferences.

Example:

All calculations are based on data in matrix K read here. To use other data, modify this section. Be sure to specify X & Y columns as done in Variable Assignment below. After that, calculations should flow properly.

K := READPRN("c:/2008LinearModelsData/Toluca.txt") **< KNNL Toluca Example p. 21**

Variable Assignment:

$X := K^{\langle 0 \rangle}$ **< X variable assignment**

$Y := K^{\langle 1 \rangle}$ **< Y variable assignment**

Summary Statistics:

n := length(Y) n = 25 **< number of observations**

X_{bar} := mean(X) $X_{bar} = 70$ **< mean of X_i**

Y_{bar} := mean(Y) $Y_{bar} = 312.28$ **< mean of Y_i**

i := 0 .. n − 1 **< range variable i**

Least Squares Estimation of the Regression Parameters:

Sums of Squares and Cross Products corrected for mean location:

$L_{xx} := \sum_i \left(X_i - X_{bar}\right)^2$ $L_{xx} = 19800$ **< Corrected Sum of Squares for X_i**

$L_{yy} := \sum_i \left(Y_i - Y_{bar}\right)^2$ $L_{yy} = 307203.04$ **< Corrected Sum of Squares for Y_i**

$L_{xy} := \sum_i \left(X_i - X_{bar}\right) \cdot \left(Y_i - Y_{bar}\right)$ $L_{xy} = 70690$ **< Corrected Sum of Squares for Cross Product**

Estimated Regression Coefficients for $Y_i = \beta_0 + \beta_1 X_i$

$b_1 := \frac{L_{xy}}{L_{xx}}$ $b_1 = 3.5702$ **< sample estimate of slope β_1**

$b_0 := Y_{bar} - b_1 \cdot X_{bar}$ $b_0 = 62.3659$ **< sample estimate of intercept β_0**

Point Estimate of the Mean response (Y_h):

$Yh_i := b_0 + b_1 \cdot X_i$ **< vector of points along the regression line.**

Residuals:

$e_i := Yh_i - Y_i$ **< vector of deviations of each value Y_i from Regression line = Yh_i**

Point Estimate of Variance of Y_h called s^2 or MSE (estimating underlying population σ^2):

$SSE := \sum_i \left(e_i\right)^2$ $SSE = 54825.4592$ **< Sum of Squares Error**

$MSE := \frac{SSE}{n-2}$ $MSE = 2383.7156$ **< Mean Squares Error**

Maximum Liklihood Estimation of the Regression Parameters:

$\beta hat_1 := b_1$ **< slope**

$\beta hat_0 := b_0$ **< intercept**

$\sigma sq_{hat} := \frac{SSE}{n}$

Coefficient of Determination (R^2) & Coefficient of Correlation (R):

$SSR := \sum_i \left(Yh_i - Y_{bar}\right)^2$ $SSR = 252377.5808$ **< Sum of Squares Regression, also in ANOVA table**

$R_{sq} := \frac{SSR}{L_{yy}}$ $R_{sq} = 0.8215$ **< Coefficient of Determination**

$R := \sqrt{R_{sq}}$ $R = 0.9064$ **< Pearson Product Moment Correlation**

$r := \frac{L_{xy}}{\sqrt{L_{xx} \cdot L_{yy}}}$ $r = 0.9064$ **< point estimate of ρ - correlation of multivariate normal population of X & Y (with no distinction between independent vs dependent). Same as R.**

Tests of Regression Parameter β_1 - Slope:

t-Test:

Hypotheses:

$\beta1_0 := 0$ **< set this as desired.**

$H_0: \beta_1 = \beta_{10}$ **< $\beta_{10} = 0$ tests for 0 slope, and No linear association between X & Y. However the test is a general one, allowing testing of any value of β_{10}**

$H_1: \beta_1 <> \beta_{10}$ < **Two sided test**

Test Statistic:

$$t := \frac{b_1 - \beta1_0}{\sqrt{\frac{MSE}{L_{xx}}}} \qquad t = 10.2896$$

< confirmed p. 47

Critical Value of the Test:

$\alpha := 0.05$ **< Probability of Type I error must be explicitly set**

$$C_1 := qt\left(\frac{\alpha}{2}, n-2\right) \quad C_1 = -2.0687 \qquad C_2 := qt\left(1 - \frac{\alpha}{2}, n-2\right) \quad C_2 = 2.0687$$

< confirmed p. 47

^ Note degrees of freedom = (n-2)

Decision Rule:

IF |t| > C, THEN REJECT H_0 OTHERWISE ACCEPT H_0

Probability Value:

$$P := \min[2 \cdot pt(t, n-2), 2 \cdot (1 - pt(t, n-2))] \qquad P = 4.4488 \times 10^{-10}$$

Confidence Interval for Regression Parameter (β_1):

$$CI_R := \left(b_1 + C_1 \cdot \sqrt{\frac{MSE}{L_{xx}}} \quad b_1 + C_2 \cdot \sqrt{\frac{MSE}{L_{xx}}}\right)$$

$CI_R = (2.8524 \quad 4.288)$ **< confirmed p. 46**

Note that C_1 and C_2 are explicitly evaluated above so C_1 is already negative in value. So it is added to b_1 here to find the Lower Bound of the CI.

ANOVA F-Test:

Sum of Squares Partition:

$$SSTO := \sum_i \left(Y_i - Y_{bar}\right)^2 \qquad L_{yy} = 307203.04 \qquad SSTO = 307203.04$$

< **Total** **Sum of Squares**

$$SSR := \sum_i \left(Yh_i - Y_{bar}\right)^2 \qquad SSR = 252377.5808$$

< **Regression** **Sum of Squares**

$$SSE := \sum_i \left(Y_i - Yh_i\right)^2 \qquad SSE = 54825.4592$$

< **Residual** **(also called "Error") Sum of Squares**

ANOVA TABLE

	SS	df	MS	
Regression:	$SSR = 252377.5808$	1	$MSR := \frac{SSR}{1}$	$MSR = 252377.5808$
Residual:	$SSE = 54825.4592$	$(n - 2)$	$MSE := \frac{SSE}{(n - 2)}$	$MSE = 2383.7156$
TOTAL:	$SSTO = 307203.04$	$(n - 1)$	$MSTO := \frac{SSTO}{(n - 1)}$	$MSTO = 12800.1267$

^ confirmed p. 71

Hypotheses:

H_0: $\beta_1 = 0$ **< No linear association between X & Y.**

H_1:β_1 <> 0 < **Two sided test**

Test Statistic:

$F := \frac{MSR}{MSE}$ $F = 105.8757$ **< confirmed p. 71**

$\sqrt{F} = 10.2896$ **< Note alternate way to obtain the t-statistic!**

Critical Value of the Test:

$\alpha := 0.05$ **< Probability of Type I error must be explicitly set**

$CV := qF(1 - \alpha, 1, n - 2)$ $CV = 4.2793$

Decision Rule:

IF F > CV, THEN REJECT H_0 OTHERWISE ACCEPT H_0

$F = 105.8757$ $CV = 4.2793$

Probability Value:

$P := 1 - pF(F, 1, n - 2)$ $P = 4.4488 \times 10^{-10}$ **< Note same value as in t-test above.**

Confidence Interval for Regression Parameter (β_1):

Calculate t-statistic directly, or convert the F-Statistic to the t-statistic, and calculate as above.

$CI_R = (2.8524 \quad 4.288)$

GLM Test Approach:

Hypotheses:

H_0: $\beta_1 = 0$ **< No linear association between X & Y.**

H_1:β_1 <> 0 < **Two sided test**

Full Model:

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ **< same model as above with all parameters β_i specified.**

Reduced Model:

$Y_i = \beta_0 + \varepsilon_i$ **< model when the null hypothesis is considered "true". Here H_0: $\beta_1 = 0$**

Error Sums of Squares for Full and Reduced Hypotheses:

$SSE_F := SSE$ $\quad SSE_F = 54825.4592$

$df_F := n - 2$ $\quad df_F = 23$

$SSE_R := SSTO$ $\quad SSE_R = 307203.04$

$df_R := n - 1$ $\quad df_R = 24$

< SS "Error" SS and df "Error" are typically reported in ANOVA tables after fitting each model separately.

Test Statistic:

$$F := \frac{\frac{SSE_R - SSE_F}{df_R - df_F}}{\frac{SSE_F}{df_F}} \qquad F = 105.8757$$

< same as F calculated above, see also p. 73

Critical Value of the Test:

$\alpha := 0.05$ **< Probability of Type I error must be explicitly set**

$CV := qF(1 - \alpha, 1, n - 2)$ $\quad CV = 4.2793$

Decision Rule:

IF F > CV, THEN REJECT H_0 OTHERWISE ACCEPT H_0

$F = 105.8757$ $\quad CV = 4.2793$

Probability Value:

$P := 1 - pF(F, 1, n - 2)$ $\quad P = 4.4488 \times 10^{-10}$ **< Same value as above.**

Tests of Regression Parameter β_0 - Intercept:

t-Test:

Hypotheses:

$\beta 0_0 := 0$ **< set this as desired.**

H_0: $\beta_0 = \beta_{00}$ **< β_{10} = 0 tests for 0 slope, and No linear association between X & Y. However the test is a general one, allowing testing of any value of β_{10}**

H_1: $\beta_0 <> \beta_{00}$ < **Two sided test**

Test Statistic:

$$t := \frac{b_0 - \beta 0_0}{\sqrt{MSE \cdot \left(\frac{1}{n} + \frac{X_{bar}^2}{L_{xx}}\right)}} \qquad t = 2.3824$$

< denominator confirmed p. 49

$$\sqrt{MSE \cdot \left(\frac{1}{n} + \frac{X_{bar}^2}{L_{xx}}\right)} = 26.1774$$

Critical Value of the Test:

$\alpha := 0.10$ **< Probability of Type I error must be explicitly set**

$C_1 := qt\left(\frac{\alpha}{2}, n - 2\right)$ $\quad C_1 = -1.7139$ $\quad C_2 := qt\left(1 - \frac{\alpha}{2}, n - 2\right)$ $\quad C_2 = 1.7139$ **< confirmed p. 49**

^ Note degrees of freedom = (n-2)

Decision Rule:

IF |t| > C, THEN REJECT H_0 OTHERWISE ACCEPT H_0

Probability Value:

$P := \min[2 \cdot pt(t, n-2), 2 \cdot (1 - pt(t, n-2))]$ $P = 0.0259$

Confidence Interval for Intercept (β_0):

$$CI_I := \left[b_0 + C_1 \cdot \sqrt{MSE \cdot \left(\frac{1}{n} + \frac{X_{bar}^2}{L_{xx}}\right)} \quad b_0 + C_2 \cdot \sqrt{MSE \cdot \left(\frac{1}{n} + \frac{X_{bar}^2}{L_{xx}}\right)} \right]$$

$CI_I = (17.5011 \quad 107.2306)$ **< confirmed p. 49**

Confidence Intervals for Regression Estimates Y_h and New Predictions of Y:

One or more values of X_n must be explicitly specified to obtain a prediction CI for Y_h:

$Xn_i := X_i$ **< here using all original values of X, but any X values may be specified instead...**

Critical Values:

$\alpha := 0.10$ **< Probability of Type I error must be explicitly set**

$C_1 := qt\left(\frac{\alpha}{2}, n-2\right)$ $C_1 = -1.7139$ $C_2 := qt\left(1 - \frac{\alpha}{2}, n-2\right)$ $C_2 = 1.7139$ **< confirmed p. 49**

^ Note degrees of freedom = (n-2)

$W := \sqrt{2 \cdot qF(1-\alpha, 2, n-2)}$ $W = 2.258$ **< confirmed p. 62**

Confidence Interval (CI):

$$CI_{RL_i} := Yh_i + C_1 \cdot \sqrt{MSE \cdot \left[\frac{1}{n} + \frac{(Xn_i - X_{bar})^2}{L_{xx}}\right]} \qquad CI_{RU_i} := Yh_i + C_2 \cdot \sqrt{MSE \cdot \left[\frac{1}{n} + \frac{(Xn_i - X_{bar})^2}{L_{xx}}\right]}$$

$\left(CI_{RL_8} \quad CI_{RU_8}\right) = (394.9251 \quad 443.847)$ **< confirmed p. 55 for point: X_8** $X_8 = 100$

Prediction Interval (PI):

$$PI_{RL_i} := Yh_i + C_1 \cdot \sqrt{MSE \cdot \left[1 + \frac{1}{n} + \frac{(Xn_i - X_{bar})^2}{L_{xx}}\right]} \qquad PI_{RU_i} := Yh_i + C_2 \cdot \sqrt{MSE \cdot \left[1 + \frac{1}{n} + \frac{(Xn_i - X_{bar})^2}{L_{xx}}\right]}$$

$\left(PI_{RL_8} \quad PI_{RU_8}\right) = (332.2072 \quad 506.5649)$ **< confirmed p. 59 for point: X_8** $X_8 = 100$

Working-Hotelling Confidence Band (WI):

Note: this is a *simulataneous* estimate for the entire regression line (Y_h), thus wider than CI.

$$WI_{RL_i} := Yh_i - W \cdot \sqrt{MSE \cdot \left[\frac{1}{n} + \frac{(Xn_i - X_{bar})^2}{L_{xx}}\right]} \qquad WI_{RU_i} := Yh_i + W \cdot \sqrt{MSE \cdot \left[\frac{1}{n} + \frac{(Xn_i - X_{bar})^2}{L_{xx}}\right]}$$

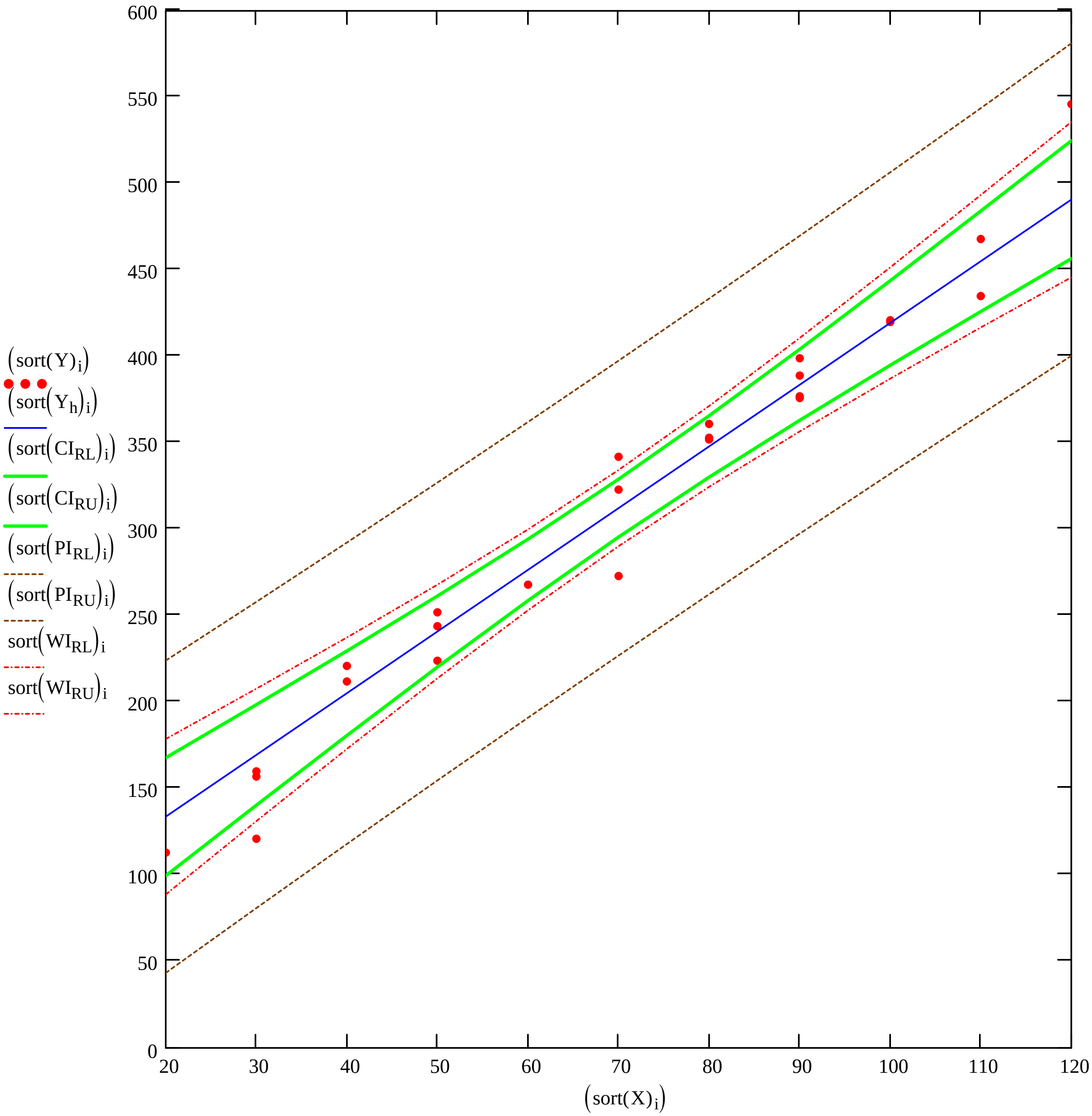
$\left(WI_{RL_8} \quad WI_{RU_8}\right) = (387.1591 \quad 451.613)$ **< confirmed p. 62 for point: X_8** $X_8 = 100$

Note: since Working-Hotelling Confidence Bands may not be available in some statistical packages, one can calculate them from Confidence Intervals (CI) above (WH):

$$WH_{RL_i} := Y_{h_i} - \left[\left(\frac{W}{C_2}\right)\cdot\left(Y_h - CI_{RL}\right)\right]_i \qquad WH_{RU_i} := Y_{h_i} + \left[\left(\frac{W}{C_2}\right)\cdot\left(Y_h - CI_{RL}\right)\right]_i$$

$\left(WH_{RL_8} \;\; WH_{RU_8}\right) = (387.1591 \;\; 451.6135)$ **same as WI above for point X_8** $\qquad X_8 = 100$

Plot of CI, PI & WI intervals:



Inference on Correlation Coefficient ρ - correlation between X & Y:

Assumptions:

- Standard Linear Regression depends on specifying in advance which variable is to be considered 'dependent' and which 'independent'. Here, it doesn't matter.

- Variables X,Y are a paired random sample drawn from a bivariate Normal distribution with marginal distributions:

- $Y_1, Y_2, Y_3, \ldots, Y_n$ (one variable) $\sim N(\mu_Y, \sigma_Y^2)$

- $X_1, X_2, X_3, \ldots, X_n$ (another variable) $\sim N(\mu_X, \sigma_X^2)$

- Let ρ be the coefficient of correlation between X & Y and estimated as above.

Hypotheses:

H_0: ρ = 0 < **β_{10} = 0 tests for 0 slope, and No linear association between X & Y. However the test is a general one, allowing testing of any value of β_{10}**

H_1: ρ <> 0 < **Two sided test**

Test Statistic:

$t := \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$ $t = 10.2896$ **< confirmed p. 84. Note same result as t-test above!**

Critical Value of the Test:

$\alpha := 0.05$ **< Probability of Type I error must be explicitly set**

$C_1 := qt\left(\frac{\alpha}{2}, n-2\right)$ $C_1 = -2.0687$ $C_2 := qt\left(1-\frac{\alpha}{2}, n-2\right)$ $C_2 = 2.0687$

^ Note degrees of freedom = (n-2)

Decision Rule:

IF |t| > C, THEN REJECT H_0 OTHERWISE ACCEPT H_0

Probability Value:

$P := \min[2 \cdot pt(t, n-2), 2 \cdot (1 - pt(t, n-2))]$ $P = 4.4488 \times 10^{-10}$

Confidence Interval for Correlation Parameter (ρ):

NOTE: n should be > 24 $n = 25$

$\xi := \frac{1}{2} \cdot \ln\left(\frac{1+r}{1-r}\right)$ $\xi = 1.5069$ **< Fisher's transformation of r into ξ**

$C_1 := qnorm\left(\frac{\alpha}{2}, 0, 1\right)$ $C_1 = -1.96$ $C_2 := qnorm\left(1-\frac{\alpha}{2}, 0, 1\right)$ $C_2 = 1.96$ **^ Note use of N(0,1) here!**

$RI_{\xi L} := \xi + C_1 \cdot \frac{1}{\sqrt{n-3}}$ $RI_{\xi U} := \xi + C_2 \cdot \frac{1}{\sqrt{n-3}}$ **< Lower and Upper bounds in ξ**

$RI_{\xi} := \left(RI_{\xi L} \quad \xi + RI_{\xi U}\right)$ $RI_{\xi} = (1.089 \quad 3.4316)$ **< interval in transformed variable ξ**

$RI_{rL} := \frac{\exp(2 \cdot RI_{\xi L}) - 1}{\exp(2 \cdot RI_{\xi L}) + 1}$ $RI_{rU} := \frac{\exp(2 \cdot RI_{\xi U}) - 1}{\exp(2 \cdot RI_{\xi U}) + 1}$ **< Lower and Upper bounds in r**

$RI_r := \left(RI_{rL} \quad RI_{rU}\right)$ $RI_r = (0.7965 \quad 0.9583)$ **< interval in original units of r**

Prototype in R:

COMMANDS:

```
#READ TABLE AND ASSIGN VARIABLES FROM COLUMNS:
K=read.table("c:/2008LinearModelsData/Toluca.txt")
attach(K)
X=V1
Y=V2

summary(lm(Y~X))
```

RESULTS:

```
Call:
lm(formula = Y ~ X)

Residuals:
   Min      1Q  Median     3Q     Max
-83.876 -34.088  -5.982  38.826 103.528

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   62.366     26.177   2.382   0.0259 *
X              3.570      0.347  10.290 4.45e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 48.82 on 23 degrees of freedom
Multiple R-Squared: 0.8215,     Adjusted R-squared: 0.8138
F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10
```

```
anova(lm(Y~X))
```

RESULTS:

```
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X          1 252378  252378  105.88 4.449e-10 ***
Residuals 23  54825    2384
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
confint(lm(Y~X),level=0.90)

(Intercept) 17.501100 107.230617
X            2.975536   4.164868

confint(lm(Y~X),level=0.95)

(Intercept) 8.213711 116.518006
X           2.852435   4.287969
```

FROM ABOVE:

estimated regression coefficients:

$$b = \begin{pmatrix} 62.3659 \\ 3.5702 \end{pmatrix}$$

t-test statistic and Probability:

$t = 10.2896$

standard error of b_1:

$$\sqrt{\frac{MSE}{L_{xx}}} = 0.347$$

standard error of b_0:

$$\sqrt{MSE \cdot \left(\frac{1}{n} + \frac{X_{bar}^{2}}{L_{xx}}\right)} = 26.1774$$

coefficient of determination & correlation:

$R_{sq} = 0.8215 \qquad r = 0.9064$

ANOVA table:

$SSR = 252377.5808 \quad MSR = 252377.5808$

$SSE = 54825.4592 \quad MSE = 2383.7156$

F-statistic and Probability:

$F = 105.8757$

Confidence Intervals on regression coefficients β_0 and β_1 at different α levels:

$CI_I = (17.5011 \quad 107.2306)$ **@ α=0.90**

$CI_R = (2.8524 \quad 4.288)$ **@ α=0.95**

predict.lm(lm(Y~X),interval="confidence",level=0.90)

RESULT:

```
   fit      lwr       upr
1  347.9820 330.22151 365.7425
2  169.4719 140.38796 198.5559
3  240.8760 220.34492 261.4070
4  383.6840 363.15300 404.2151
5  312.2800 295.54462 329.0154
6  276.5780 258.81747 294.3385
7  490.7901 456.67059 524.9096
8  347.9820 330.22151 365.7425
9  419.3861 394.92512 443.8470
10 240.8760 220.34492 261.4070
11 205.1739 180.71300 229.6349
12 312.2800 295.54462 329.0154
13 383.6840 363.15300 404.2151
14 133.7699  99.65039 167.8894
15 455.0881 426.00412 484.1720
16 419.3861 394.92512 443.8470
17 169.4719 140.38796 198.5559
18 240.8760 220.34492 261.4070
19 383.6840 363.15300 404.2151
20 455.0881 426.00412 484.1720
21 169.4719 140.38796 198.5559
22 383.6840 363.15300 404.2151
23 205.1739 180.71300 229.6349
24 347.9820 330.22151 365.7425
25 312.2800 295.54462 329.0154
```

Confidence Interval on Y_h:

$CI := \text{augment}(Y_h, CI_{RL}, CI_{RU})$

CI =

	0	1	2
0	347.982	330.2215	365.7425
1	169.4719	140.388	198.5559
2	240.876	220.3449	261.407
3	383.684	363.153	404.2151
4	312.28	295.5446	329.0154
5	276.578	258.8175	294.3385
6	490.7901	456.6706	524.9096
7	347.982	330.2215	365.7425
8	419.3861	394.9251	443.847
9	240.876	220.3449	261.407
10	205.1739	180.713	229.6349
11	312.28	295.5446	329.0154
12	383.684	363.153	404.2151
13	133.7699	99.6504	167.8894
14	455.0881	426.0041	484.172
15	419.3861	394.9251	443.847
16	169.4719	140.388	198.5559
17	240.876	220.3449	261.407
18	383.684	363.153	404.2151
19	455.0881	426.0041	484.172
20	169.4719	140.388	198.5559
21	383.684	363.153	404.2151
22	205.1739	180.713	229.6349
23	347.982	330.2215	365.7425
24	312.28	295.5446	329.0154

predict.lm(lm(Y~X),interval="prediction",level=0.90)

RESULT:

```
      fit       lwr       upr
1  347.9820 262.44106 433.5230
2  169.4719  80.88469 258.0591
3  240.8760 154.71713 327.0348
4  383.6840 297.52521 469.8429
5  312.2800 226.94599 397.6140
6  276.5780 191.03702 362.1189
7  490.7901 400.42439 581.1558
8  347.9820 262.44106 433.5230
9  419.3861 332.20718 506.5649
10 240.8760 154.71713 327.0348
11 205.1739 117.99506 292.3528
12 312.2800 226.94599 397.6140
13 383.6840 297.52521 469.8429
14 133.7699  43.40419 224.1356
15 455.0881 366.50085 543.6753
16 419.3861 332.20718 506.5649
17 169.4719  80.88469 258.0591
18 240.8760 154.71713 327.0348
19 383.6840 297.52521 469.8429
20 455.0881 366.50085 543.6753
21 169.4719  80.88469 258.0591
22 383.6840 297.52521 469.8429
23 205.1739 117.99506 292.3528
24 347.9820 262.44106 433.5230
25 312.2800 226.94599 397.6140
Warning message:
Predictions on current data refer to
_future_ responses
 in: predict.lm(lm(Y ~ X), interval =
"prediction", level = 0.9)
```

^ Yes, as requested

Prediction Interval on Y_h:

$PI := \text{augment}\left(Y_h, PI_{RL}, PI_{RU}\right)$

PI =

	0	1	2
0	347.982	262.4411	433.523
1	169.4719	80.8847	258.0591
2	240.876	154.7171	327.0348
3	383.684	297.5252	469.8429
4	312.28	226.946	397.614
5	276.578	191.037	362.1189
6	490.7901	400.4244	581.1558
7	347.982	262.4411	433.523
8	419.3861	332.2072	506.5649
9	240.876	154.7171	327.0348
10	205.1739	117.9951	292.3528
11	312.28	226.946	397.614
12	383.684	297.5252	469.8429
13	133.7699	43.4042	224.1356
14	455.0881	366.5009	543.6753
15	419.3861	332.2072	506.5649
16	169.4719	80.8847	258.0591
17	240.876	154.7171	327.0348
18	383.684	297.5252	469.8429
19	455.0881	366.5009	543.6753
20	169.4719	80.8847	258.0591
21	383.684	297.5252	469.8429
22	205.1739	117.9951	292.3528
23	347.982	262.4411	433.523
24	312.28	226.946	397.614

cor.test(X,Y, method="pearson",alternative="two.sided",conf.level=0.95)

RESULT:

```
        Pearson's product-moment correlation

data:  X and Y
t = 10.2896, df = 23, p-value = 4.449e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7965202 0.9583070
sample estimates:
      cor
0.9063848
```

correlation t-test is identical with other t-tests above:

$t = 10.2896$

Fisher's z-tranform Confidence Interval for ρ is confirmed in R with Toluca data.

$RI_r = (\,0.7965202 \quad 0.958307\,)$

correlation reported is the same:

$r = 0.90638484$

```
#COLLECTING INFORMATION FROM THE REGRESSION:
Yh=predict(lm(Y~X))
RESID=resid(lm(Y~X))
CONF=data.frame(predict.lm(lm(Y~X),interval="confidence",level=0.90))
PRED= data.frame(predict.lm(lm(Y~X),interval="prediction",level=0.90))
#CALCULATING WORKING-HOTELLING INTERVAL ALSO:
#FIRST ONE SPECIFIES CONFIDENCE LEVEL ALPHA:
alpha=0.10
#FINDING n:
n=length(X)
#USING qt() FUNCTION TO CALCULATE CRITICAL VALUE C :
C=qt((1-(alpha/2)),(n-2))
C
#USING qf() FUNCTION TO CALCULATE WI CRITICAL VALUE W:
W=sqrt(2*qf(1-alpha,2,(n-2)))
W
#CALCULATING WORKING-HOTELLING INTERVALS:
WH.lwr=Yh-((W/C)*(Yh-CONF$lwr))
WH.upr=Yh+((W/C)*(Yh-CONF$lwr))
WORHOT=data.frame(WH.lwr,WH.upr)
WORHOT
# REPORTING WITHIN A DATAFRAME CALLED "RESULTS":
RESULTS=data.frame(X,Y,Yh,RESID,CONF$lwr,CONF$upr,PRED$lwr,PRED$upr,WH.lwr,WH.upr)
RESULTS
```

RESULT:

	X	Y	Yh	RESID	CONF.lwr	CONF.upr	PRED.lwr	PRED.upr	WH.lwr	WH.upr
1	80	399	347.9820	51.0179798	330.22151	365.7425	262.44106	433.5230	324.58279	371.3813
2	30	121	169.4719	-48.4719192	140.38796	198.5559	80.88469	258.0591	131.15419	207.7897
3	50	221	240.8760	-19.8759596	220.34492	261.4070	154.71713	327.0348	213.82658	267.9253
4	90	376	383.6840	-7.6840404	363.15300	404.2151	297.52521	469.8429	356.63466	410.7334
5	70	361	312.2800	48.7200000	295.54462	329.0154	226.94599	397.6140	290.23136	334.3286
6	60	224	276.5780	-52.5779798	258.81747	294.3385	191.03702	362.1189	253.17875	299.9772
7	120	546	490.7901	55.2098990	456.67059	524.9096	400.42439	581.1558	445.83809	535.7421
8	80	352	347.9820	4.0179798	330.22151	365.7425	262.44106	433.5230	324.58279	371.3813
9	100	353	419.3861	-66.3860606	394.92512	443.8470	332.20718	506.5649	387.15910	451.6130
10	50	157	240.8760	-83.8759596	220.34492	261.4070	154.71713	327.0348	213.82658	267.9253
11	40	160	205.1739	-45.1739394	180.71300	229.6349	117.99506	292.3528	172.94698	237.4009
12	70	252	312.2800	-60.2800000	295.54462	329.0154	226.94599	397.6140	290.23136	334.3286
13	90	389	383.6840	5.3159596	363.15300	404.2151	297.52521	469.8429	356.63466	410.7334
14	20	113	133.7699	-20.7698990	99.65039	167.8894	43.40419	224.1356	88.81789	178.7219
15	110	435	455.0881	-20.0880808	426.00412	484.1720	366.50085	543.6753	416.77035	493.4058
16	100	420	419.3861	0.6139394	394.92512	443.8470	332.20718	506.5649	387.15910	451.6130
17	30	212	169.4719	42.5280808	140.38796	198.5559	80.88469	258.0591	131.15419	207.7897
18	50	268	240.8760	27.1240404	220.34492	261.4070	154.71713	327.0348	213.82658	267.9253
19	90	377	383.6840	-6.6840404	363.15300	404.2151	297.52521	469.8429	356.63466	410.7334
20	110	421	455.0881	-34.0880808	426.00412	484.1720	366.50085	543.6753	416.77035	493.4058
21	30	273	169.4719	103.5280808	140.38796	198.5559	80.88469	258.0591	131.15419	207.7897
22	90	468	383.6840	84.3159596	363.15300	404.2151	297.52521	469.8429	356.63466	410.7334
23	40	244	205.1739	38.8260606	180.71300	229.6349	117.99506	292.3528	172.94698	237.4009
24	80	342	347.9820	-5.9820202	330.22151	365.7425	262.44106	433.5230	324.58279	371.3813
25	70	323	312.2800	10.7200000	295.54462	329.0154	226.94599	397.6140	290.23136	334.3286

^ values confirmed above

#PLOTTING DATA, REGRESSION LINE, AND INTERVALS:
plot(X,Y,pch=19,col="black")
points(X,Yh,pch=19,col="blue")
abline(lm(Y~X),col="blue")
#segments(X,predict(lm(Y~X)),X,Y,col="red")
points(X,CONF$lwr, pch=24,col="green")
points(X,CONF$upr, pch=25,col="green")
points(X,PRED$lwr, pch=24,col="magenta")
points(X,PRED$upr, pch=25,col="magenta")
points(X,WH.lwr, pch=24,col="brown")
points(X,WH.upr, pch=25,col="brown")

RESULT: