ORIGIN ≡ 1

ANOVA F-test/t-test for Simple Linear Regression and Interval Estimation

Goodness of fit of a fitted regression line can be tested using the F-test for Regression (also known as the ANOVA for Regression). Alternatively, and equivalently, fit of the regression can be tested using a t-test approach. The latter method also allows estimation of confidence intervals for the slope parameter β. Interval estimates can also be derived for the Regression line itself as a *mean,* as well as for *prediction* of "new" observations.

ZAR := READPRN("c:/DATA/Biostatistics/ZarEX17.1R.txt")

$X := ZAR^{\langle 2\rangle}$ **^Zar Example 17.1**

$Y := ZAR^{\langle 3\rangle}$

$X_{bar} := \text{mean}(X)$ $X_{bar} = 10$

$Y_{bar} := \text{mean}(Y)$ $Y_{bar} = 3.4154$

$s := \sqrt{\text{Var}(Y)}$ $s = 1.2799$ $s^2 = 1.6381$

$n := \text{length}(Y)$ $n = 13$

$i := 1 .. n$

$X = (3, 4, 5, 6, 8, 9, 10, 11, 12, 14, 15, 16, 17)^T$

$Y = (1.4, 1.5, 2.2, 2.4, 3.1, 3.2, 3.2, 3.9, 4.1, 4.7, 4.5, 5.2, 5)^T$

ZAR =

	1	2	3
1	1	3	1.4
2	2	4	1.5
3	3	5	2.2
4	4	6	2.4
5	5	8	3.1
6	6	9	3.2
7	7	10	3.2
8	8	11	3.9
9	9	12	4.1
10	10	14	4.7
11	11	15	4.5
12	12	16	5.2
13	13	17	5

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.**
- **$Y_1, Y_2, Y_3, \ldots, Y_n$ (dependent variable) is a random sample ~ $N(\mu,\sigma^2)$.**
- **$X_1, X_2, X_3, \ldots, X_n$ (independent variable) with each value of X_i matched to Y_i**

Model:

$Y = \alpha + \beta X + \varepsilon$

where: α is the y intercept of the regression line (translation)
β is the slope of the regression line (scaling coefficient)
ε is the error factor in prediction of Y given that it is a random variable with $N(0,\sigma^2)$

Corrected Sums of Squares:

$L_{xx} := \sum_i (X_i - X_{bar})^2$ $L_{yy} := \sum_i (Y_i - Y_{bar})^2$ $L_{xy} := \sum_i (X_i - X_{bar})\cdot(Y_i - Y_{bar})$

$L_{xx} = 262$ $L_{yy} = 19.6569$ $L_{xy} = 70.8$

Estimated Regression Coefficients for Y = α + βX:

$b := \frac{L_{xy}}{L_{xx}}$ $b = 0.2702$ **< sample estimate of β**

$a := Y_{bar} - b\cdot X_{bar}$ $a = 0.7131$ **< sample estimate of α**

Estimated values of Y (Y_{hat}):

$Y_{hat_i} := a + b\cdot X_i$

Residuals:

$e_i := Y_{hat_i} - Y_i$

Sum of Squares Decomposition of the Regression:

Once a regression model ($Y = \alpha + \beta X + \varepsilon$) is fitted with data as $Y_{hat} = a + bX$, one still needs to determine how useful the regression might be, especially whether knowledge about the X_i provide insight into interpreting the Y_i as a random variable from a Normal distribution with error ε_i.
This is done by considering a "partition" of total Sums of Squares in the sample of Y_i.

$SS_T := \sum_i \left(Y_i - Y_{bar}\right)^2$ $SS_T = 19.6569$ $L_{yy} = 19.6569$ < **Total Sum of Squares**

$SS_R := \sum_i \left(Y_{hat_i} - Y_{bar}\right)^2$ $SS_R = 19.1322$ < **Regression Sum of Squares**

$SS_E := \sum_i \left(Y_i - Y_{hat_i}\right)^2$ $SS_E = 0.5247$ < **Residual (also called "Error") Sum of Squares**

These Sums of Squares tally as follows:

$SS_T = 19.6569$ $SS_R + SS_E = 19.6569$

And the ratio of SS_R to SS_E can be used as a measure of "fit" of the data to the regression.

ANOVA Table for Linear Regression:

	SS	df	MS	
Regression:	$SS_R = 19.1322$	$df_R := 1$	$MS_R := \frac{SS_R}{df_R}$	$MS_R = 19.1322$
Residual:	$SS_E = 0.5247$	$df_E := n - 2$	$MS_E := \frac{SS_E}{df_E}$	$MS_E = 0.0477$
TOTAL:	$SS_T = 19.6569$	$df_T := n - 1$	$MS_T := \frac{SS_T}{df_T}$	$MS_T = 1.6381$

ANOVA F-test for Regression Slope:

Hypotheses:

H_0: $\beta = 0$ < **Slope of the Regression is zero implying no relationship between X_i and Y_i**
H_1: $\beta \neq 0$ < **Two sided test**

Test Statistic:

$F := \frac{MS_R}{MS_E}$ $F = 401.0875$ < **F is the ratio of sample variances**

Sampling Distribution:

If Assumptions hold and H_0 is true, then $F_s \sim F_{(1,n-2)}$

Critical Value of the Test:

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

$C := qF(1 - \alpha, 1, n - 2)$ $C = 4.8443$

Decision Rule:

IF F > C, THEN REJECT H_0, OTHERWISE ACCEPT H_0

Probability Value:

$$P := 1 - pF(F, 1, n-2) \qquad P = 5.2671 \times 10^{-10}$$

Prototype in R:

```
#SIMPLE LINEAR REGRESSION
#ZAR EXAMPLE 17.1
ZAR=read.table("c:/DATA/Biostatistics/ZarEX17.1R.txt")
ZAR
attach(ZAR)
X=ageX
Y=winglY

#CREATING LINEAR MODEL LM:
LM=lm(Y~X)
Yhat=fitted(LM)
e=residuals(LM)
RESULTS=data.frame(X,Y,Yhat,e)
RESULTS

#ANOVA F-TEST:
anova(LM)
```

```
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
X          1 19.1322 19.1322  401.09 5.267e-10 ***
Residuals 11  0.5247  0.0477
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The t-Test Approach and Interval Estimation for Simple Linear Regression:

The t-test approach is exactly equivalent to the F-test above with the statistic F = statistic t^2. In contrast to the F-test, however, the t-test allows for one-tailed tests and for construction of confidence intervals of the regression (to capture β), confidence intervals for the regression estimates (to capture Y_{hat}) and prediction intervals for new observations (to capture new observations X_n).

t-Test for Regression Parameter β (slope):

Hypotheses:

$H_0: \beta = 0$ **< Slope of the Regression is zero implying no relationship between X_i and Y_i**

$H_1: \beta \neq 0$ **< Two Sided Case**

$H_1: \beta < 0$ **< One Sided Case Lower Tail)**

$H_1: \beta > 0$ **< One Sided Case (Upper Tail)**

Test Statistic:

$$t_\beta := \frac{b}{\sqrt{\frac{MSE}{L_{xx}}}} \qquad t_\beta = 20.0272$$

< b is unbiased point estimate of β

< MSE is Mean Square Error from ANOVA table also denoted $s^2_{x.y}$

< L_{xx} Corrected sums of squares of X as defined in Regression

Sampling Distribution of Test Statistic t_β:

If Assumptions hold and H_0 is true, then $t_\beta \sim t_{(n-2)}$

Critical Values of the Test:

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

$C_1 := qt\left(\frac{\alpha}{2}, n-2\right)$ $C_1 = -2.201$ **< Two sided lower Critical Value**

$C_2 := qt\left(1-\frac{\alpha}{2}, n-2\right)$ $C_2 = 2.201$ **< Two sided upper Critical Value**

$C := |C_2|$ $C = 2.201$ **< Critical value used for two sided test (to simplify)**

$C_3 := qt(\alpha, n-2)$ $C_3 = -1.7959$ **< One sided lower Critical Value**

$C_4 := qt(1-\alpha, n-2)$ $C_4 = 1.7959$ **< One sided upper Critical Value**

Decision Rules:

IF $|t_\beta| > C$ THEN REJECT H_0, OTHERWISE ACCEPT H_0 **< Two sided case**
IF $t_\beta < C_3$, THEN REJECT H_0, OTHERWISE ACCEPT H_0 **< One sided case lower tail**
IF $t_\beta > C_4$, THEN REJECT H_0, OTHERWISE ACCEPT H_0 **< One sided case upper tail**

Probability Values:

$P_\beta := 2 \cdot pt(t_\beta, n-2)$ $P_\beta = 2$ **< if** $t \le 0$

$P_\beta := 2 \cdot (1 - pt(t_\beta, n-2))$ $P_\beta = 5.267053 \times 10^{-10}$ **< if $t > 0$** **< Two sided case**

$P_\beta := pt(t_\beta, n-2)$ $P_\beta = 1$ **< One sided case lower tail**

$P_\beta := 1 - pt(t_\beta, n-2)$ $P_\beta = 2.633527 \times 10^{-10}$ **< One sided case upper tail**

Confidence Interval for the Regression Slope (β):

$CI := \left(b - C \cdot \sqrt{\frac{MS_E}{L_{xx}}} \quad b + C \cdot \sqrt{\frac{MS_E}{L_{xx}}}\right)$ $CI = (0.2405308 \quad 0.2999272)$ **< Two sided case** $b = 0.2702$

$CIL := b - C_3 \cdot \sqrt{\frac{MS_E}{L_{xx}}}$ **minus infinity to** $CIL = 0.294461$ **< One sided case lower tail**

$CIU := b - C_4 \cdot \sqrt{\frac{MS_E}{L_{xx}}}$ $CIU = 0.245997$ **to infinity** **< One sided case upper tail**

t-Test for Regression Parameter α (intercept):

Hypotheses:

$\alpha_0 := 0$ **< any value may be tested**

H_0: $\alpha = \alpha_0$ **< tests for intercept = α_0=0 here, but may be set to other values as desired**

H_1: $\alpha \neq \alpha_0$ **< Two Sided Case**

H_1: $\alpha_0 < 0$ **< One Sided Case Lower Tail)**

H_1: $\alpha_0 > 0$ **< One Sided Case (Upper Tail)**

Test Statistic:

$$t_\alpha := \frac{a - \alpha_0}{\sqrt{MS_E \cdot \left(\frac{1}{n} + \frac{X_{bar}^2}{L_{xx}}\right)}} \qquad t_\alpha = 4.8213$$

Sampling Distribution of Test Statistic t_α:

If Assumptions hold and H_0 is true, then $t_\alpha \sim t_{(n-2)}$

Critical Values of the Test:

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

$C_1 := qt\left(\frac{\alpha}{2}, n-2\right)$ $\qquad C_1 = -2.201$

$C_2 := qt\left(1 - \frac{\alpha}{2}, n-2\right)$ $\qquad C_2 = 2.201$ **< Note degrees of freedom = (n-2)**

$C := |C_1|$ $\qquad C = 2.201$

Decision Rules:

IF $|t_\alpha| > C$ THEN REJECT H_0, OTHERWISE ACCEPT H_0 **< Two sided case**
IF $t_\alpha < C_3$, THEN REJECT H_0, OTHERWISE ACCEPT H_0 **< One sided case lower tail**
IF $t_\alpha > C_4$, THEN REJECT H_0, OTHERWISE ACCEPT H_0 **< One sided case upper tail**

Probability Values:

$P_\alpha := 2 \cdot pt(t_\alpha, n-2)$ $\qquad P_\alpha = 1.9995$ **< if** $t \leq 0$

$P_\alpha := 2 \cdot (1 - pt(t_\alpha, n-2))$ $\qquad P_\alpha = 0.000535$ **< if t > 0** **< Two sided case**

$P_\alpha := pt(t_\alpha, n-2)$ $\qquad P_\alpha = 0.999733$ **< One sided case lower tail**

$P_\alpha := 1 - pt(t_\alpha, n-2)$ $\qquad P = 5.267053 \times 10^{-10}$ **< One sided case upper tail**

Confidence Interval for Intercept (α):

$a = 0.7131$

$$CI := \left[a - C \cdot \sqrt{MS_E \cdot \left(\frac{1}{n} + \frac{X_{bar}^2}{L_{xx}} \right)} \quad a + C \cdot \sqrt{MS_E \cdot \left(\frac{1}{n} + \frac{X_{bar}^2}{L_{xx}} \right)} \right]$$

$CI = (0.387559 \quad 1.03863)$ **< Two sided case**

$$CIL := a - C_3 \cdot \sqrt{MS_E \cdot \left(\frac{1}{n} + \frac{X_{bar}^2}{L_{xx}} \right)}$$

minus infinity to $CIL = 0.978714$ **< One sided case lower tail**

$$CIU := a - C_4 \cdot \sqrt{MS_E \cdot \left(\frac{1}{n} + \frac{X_{bar}^2}{L_{xx}} \right)}$$

$CIU = 0.447475$ **to infinity** **< One sided case upper tail**

Prototype in R:

```
#t-TESTS FOR SLOPE & INTERCEPT:
summary(LM)
```

$F = 401.0875$

$t_\beta = 20.0272 \qquad t_\alpha = 4.8213$

$t_\beta^2 = 401.0875$

^ Note for test of β, (slope):
Statistics $F = t_\beta^2$

```
> summary(LM)
Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.30699 -0.21538  0.06553  0.16324  0.22507 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.71309    0.14790   4.821 0.000535 ***
X                0.27023    0.01349  20.027 5.27e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2184 on 11 degrees of freedom
Multiple R-squared: 0.9733,     Adjusted R-squared: 0.9709 
F-statistic: 401.1 on 1 and 11 DF,  p-value: 5.267e-10
```

```
#CONFIDENCE INTERVAL FOR SLOPE:
#CALCULATED BY HAND!
n=length(Y)
Lxx=sum((X-mean(X))^2)
Lxx
Lxy=sum((X-mean(X))*(Y-mean(Y)))
Lxy
b=Lxy/Lxx
b
MSE=summary(LM)$sigma^2
MSE
alpha=0.05
C=abs(qt(1-alpha/2,n-2))
C
stderr=sqrt(MSE/Lxx)
CIL=b-C*stderr
CIL
CIU=b+C*stderr
CIU
CONFIDENCE INTERVALS FOR
#REGRESSION PARAMETERS
#THE EASIER WAY:
confint(lm(Y~X),level=0.95)
```

```
> Lxx
[1] 262
> Lxy
[1] 70.8
> MSE
[1] 0.04770085
> CIL
[1] 0.2405308
> CIU
[1] 0.2999272
```

```
> confint(lm(Y~X),level=0.95)
                 2.5 %    97.5 %
(Intercept)  0.3875590 1.0386300
X            0.2405308 0.2999272
```

Confidence Interval for Regression Estimates Y_{hat} and New Predictions of Y:

One or more values of X_n must be explicitly specified to obtain a prediction CI for Y_{hat}:

$$i := 1..n$$

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

Confidence Interval for Regression (CI):

$$CI := \left[Y_{hat} - C \cdot \sqrt{MS_E \cdot \left[\frac{1}{n} + \frac{(Xn - X_{bar})^2}{L_{xx}} \right]} \quad Y_{hat} + C \cdot \sqrt{MS_E \cdot \left[\frac{1}{n} + \frac{(Xn - X_{bar})^2}{L_{xx}} \right]} \right]$$

< Two sided case

$$CIL := Y_{hat} - C_3 \cdot \sqrt{MS_E \cdot \left[\frac{1}{n} + \frac{(Xn - X_{bar})^2}{L_{xx}} \right]}$$

< One sided case lower tail

$$CIU := Y_{hat} - C_4 \cdot \sqrt{MS_E \cdot \left[\frac{1}{n} + \frac{(Xn - X_{bar})^2}{L_{xx}} \right]}$$

< One sided case upper tail

minus infinity to

$$Y_{hat} = \begin{pmatrix} 1.5238 \\ 1.794 \\ 2.0642 \\ 2.3345 \\ 2.8749 \\ 3.1452 \\ 3.4154 \\ 3.6856 \\ 3.9558 \\ 4.4963 \\ 4.7665 \\ 5.0368 \\ 5.307 \end{pmatrix} \quad CI = \left[\begin{pmatrix} 1.2768 \\ 1.5715 \\ 1.8647 \\ 2.1559 \\ 2.729 \\ 3.0086 \\ 3.2821 \\ 3.549 \\ 3.8099 \\ 4.3177 \\ 4.567 \\ 4.8142 \\ 5.06 \end{pmatrix} \begin{pmatrix} 1.7707 \\ 2.0166 \\ 2.2638 \\ 2.513 \\ 3.0209 \\ 3.2817 \\ 3.5487 \\ 3.8222 \\ 4.1018 \\ 4.6749 \\ 4.9661 \\ 5.2593 \\ 5.554 \end{pmatrix} \right] \quad CIL = \begin{pmatrix} 1.725293 \\ 1.975596 \\ 2.227071 \\ 2.480171 \\ 2.994019 \\ 3.256607 \\ 3.52417 \\ 3.797065 \\ 4.074935 \\ 4.642004 \\ 4.929361 \\ 5.218344 \\ 5.508499 \end{pmatrix} \quad CIU = \begin{pmatrix} 1.32227 \\ 1.612425 \\ 1.901408 \\ 2.188766 \\ 2.755834 \\ 3.033704 \\ 3.306599 \\ 3.574162 \\ 3.83675 \\ 4.350598 \\ 4.603698 \\ 4.855173 \\ 5.105477 \end{pmatrix}$$

to infinity

Prototype in R:

```
#CONFIDENCE INTERVALS FOR Yhat:
CONF=predict(lm(Y~X),interval="confidence",level=0.95)
CN=data.frame(CONF)
CN
```

```
> CN
        fit      lwr      upr
1  1.523782 1.276815 1.770748
2  1.794011 1.571465 2.016556
3  2.064240 1.864678 2.263801
4  2.334469 2.155899 2.513038
5  2.874927 2.728970 3.020883
6  3.145156 3.008564 3.281747
7  3.415385 3.282061 3.548709
8  3.685614 3.549022 3.822205
9  3.955843 3.809886 4.101799
10 4.496301 4.317731 4.674870
11 4.766530 4.566968 4.966091
12 5.036759 4.814213 5.259304
13 5.306988 5.060021 5.553954
```

Prediction Interval for Regression (PI):

$$PI := \left[Y_{hat} - C \cdot \sqrt{MS_E \cdot \left[1 + \frac{1}{n} + \frac{(Xn - X_{bar})^2}{L_{xx}}\right]} \quad Y_{hat} + C \cdot \sqrt{MS_E \cdot \left[1 + \frac{1}{n} + \frac{(Xn - X_{bar})^2}{L_{xx}}\right]} \right]$$

< **Two sided case**

$$PIL := Y_{hat} - C_3 \cdot \sqrt{MS_E \cdot \left[1 + \frac{1}{n} + \frac{(Xn - X_{bar})^2}{L_{xx}}\right]}$$

< **One sided case lower tail**

$$PIU := Y_{hat} - C_4 \cdot \sqrt{MS_E \cdot \left[1 + \frac{1}{n} + \frac{(Xn - X_{bar})^2}{L_{xx}}\right]}$$

< **One sided case upper tail**

minus infinity to

$$Y_{hat} = \begin{pmatrix} 1.5238 \\ 1.794 \\ 2.0642 \\ 2.3345 \\ 2.8749 \\ 3.1452 \\ 3.4154 \\ 3.6856 \\ 3.9558 \\ 4.4963 \\ 4.7665 \\ 5.0368 \\ 5.307 \end{pmatrix} \quad PI = \left[\begin{pmatrix} 0.9833 \\ 1.2643 \\ 1.5438 \\ 1.8217 \\ 2.3726 \\ 2.6454 \\ 2.9165 \\ 3.1859 \\ 3.4535 \\ 3.9835 \\ 4.246 \\ 4.507 \\ 4.7666 \end{pmatrix} \begin{pmatrix} 2.0642 \\ 2.3237 \\ 2.5847 \\ 2.8473 \\ 3.3773 \\ 3.6449 \\ 3.9142 \\ 4.1853 \\ 4.4582 \\ 5.0091 \\ 5.287 \\ 5.5665 \\ 5.8474 \end{pmatrix} \right] \quad PIL = \begin{pmatrix} 1.964748 \\ 2.226235 \\ 2.488926 \\ 2.752887 \\ 3.284839 \\ 3.552913 \\ 3.822422 \\ 4.093371 \\ 4.365755 \\ 4.914719 \\ 5.191217 \\ 5.468983 \\ 5.747954 \end{pmatrix} \quad PIU = \begin{pmatrix} 1.082815 \\ 1.361786 \\ 1.639553 \\ 1.91605 \\ 2.465015 \\ 2.737398 \\ 3.008348 \\ 3.277856 \\ 3.545931 \\ 4.077882 \\ 4.341843 \\ 4.604534 \\ 4.866021 \end{pmatrix}$$

to infinity

Prototype in R:

```
#PREDICTION INTERVALS FOR Xn:
PRED=predict(lm(Y~X),interval="prediction",level=0.95)
PR=data.frame(PRED)
PR
```

```
> PR
        fit       lwr      upr
1  1.523782 0.9833454 2.064218
2  1.794011 1.2642884 2.323733
3  2.064240 1.5437554 2.584724
4  2.334469 1.8216666 2.847271
5  2.874927 2.3725501 3.377303
6  3.145156 2.6454195 3.644892
7  3.415385 2.9165317 3.914238
8  3.685614 3.1858775 4.185350
9  3.955843 3.4534661 4.458219
10 4.496301 3.9834986 5.009103
11 4.766530 4.2460455 5.287014
12 5.036759 4.5070365 5.566481
13 5.306988 4.7665515 5.847424
```

```
> PRED=predict(lm(Y~X),interval="prediction",level=0.95)
Warning message:
In predict.lm(lm(Y ~ X), interval = "prediction", level =
0.95) :
  Predictions on current data refer to _future_ responses
```

```
#PLOTTING INTERVALS IN R:
plot(X,Y)
abline(lm(Y~X),col="blue")
segments(X,PR$lwr,X,PR$upr,col="red")
segments(X,CN$lwr,X,CN$upr,col="green")
points(X,CN$lwr,col="green")
points(X,CN$upr,col="green")
points(X,PR$lwr,col="red")
points(X,PR$upr,col="red")
points(X,PR$fit,col="blue")
```