ORIGIN ≡ 0

# Multiple Regression

**Multiple Regression is an extension of the technique of linear regression to describe the relationship between a single dependent (response) variable (Y) and multiple independent (predictor) variables ($X_1$, $X_2$, $X_3$, ...). Typically, multiple regression involves specifying one, or sometimes several, linear models, constructing the multiple regression, and then testing hypotheses often involving several regression coefficients ($\beta_1$, $\beta_2$, $\beta_3$, ...) corresponding to each of the X variables.**

ZAR := READPRN("c:/DATA/Biostatistics/ZarEX20.1aR.txt" )

$Y := ZAR^{\langle 5 \rangle}$ **< dependent variable Y**

$X1 := ZAR^{\langle 1 \rangle}$ **< independent variables X1-X4**

$X2 := ZAR^{\langle 2 \rangle}$ $X3 := ZAR^{\langle 3 \rangle}$ $X4 := ZAR^{\langle 4 \rangle}$

n := length(Y) n = 33

k := 5 **< number of independent variables +1**

i := 0 .. n − 1
ii := 0 .. n − 1 **< range variables i, ii for n observations**

j := 0 .. k − 1 **< range variable j for columns of X**

ZAR =

| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 0 | 1 | 6 | 9.9 | 5.7 | 1.6 | 2.12 |
| 1 | 2 | 1 | 9.3 | 6.4 | 3 | 3.39 |
| 2 | 3 | -2 | 9.4 | 5.7 | 3.4 | 3.61 |
| 3 | 4 | 11 | 9.1 | 6.1 | 3.4 | 1.72 |
| 4 | 5 | -1 | 6.9 | 6 | 3 | 1.8 |
| 5 | 6 | 2 | 9.3 | 5.7 | 4.4 | 3.21 |
| 6 | 7 | 5 | 7.9 | 5.9 | 2.2 | 2.59 |
| 7 | 8 | 1 | 7.4 | 6.2 | 2.2 | 3.25 |
| 8 | 9 | 1 | 7.3 | 5.5 | 1.9 | 2.86 |
| 9 | 10 | 3 | 8.8 | 5.2 | 0.2 | 2.32 |
| 10 | 11 | 11 | 9.8 | 5.7 | 4.2 | 1.57 |
| 11 | 12 | 9 | 10.5 | 6.1 | 2.4 | 1.5 |
| 12 | 13 | 5 | 9.1 | 6.4 | 3.4 | 2.69 |
| 13 | 14 | -3 | 10.1 | 5.5 | 3 | 4.06 |
| 14 | 15 | 1 | 7.2 | 5.5 | 0.2 | 1.98 |

## Assumptions:

- **Multiple Linear Regression depends on specifying in advance which variable is considered 'dependent' and which others 'independent'. This decision matters as changing roles for Y versus X's usually produces a different result.**
- **$Y_1$, $Y_2$, $Y_3$, ... , $Y_n$ (dependent variable) is a random sample ~ $N(\mu,\sigma^2)$.**

- **k Vectors of *fixed* Independent Variables:**

- **$X_{1,1}$, $X_{1,2}$, $X_{1,3}$, ... , $X_{1,n}$ (independent variable) with each value of $X_{1,i}$ matched to $Y_i$**
- **$X_{2,1}$, $X_{2,2}$, $X_{2,3}$, ... , $X_{2,n}$ (independent variable) with each value of $X_{2,i}$ matched to $Y_i$**

...
- **$X_{k,1}$, $X_{k,2}$, $X_{k,3}$, ... , $X_{k,n}$ (independent variable) with each value of $Xk_{,i}$ matched to $Y_i$**

## Model:

$$Y_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \dots + \beta_k X_{k,i} + \varepsilon_i \quad \text{for } i = 1 \text{ to } n$$

**where: $\alpha$ is the y intercept of the regression line (translation).**
**$\beta_i$'s are the regression coefficient (i.e., "slope") for each $X_i$ of the regression line.**
**$\varepsilon_i$'s are the residuals (i.e. "error") in prediction of Y given that it is a random variable with $N(0,\sigma^2)$**

**Note that this is one of many possible linear models involving $X_i$'s that may be squared or higher order functions of an original variable (i.e., $X^2$, $X^3$, etc.) or cross products of two original variables (i.e., $X_{a,i}X_{b,i}$ etc.). This is the wonderful and extremely powerful world of Linear Modeling.**

## Least Squares Estimation of the Regression Line:

**Calculations in Multiple Regression are extensive, and best visualized using matrix algebra where sums of squares and cross products are implicit in matrix manipulations:**

- **X:** becomes a (n X k+1) Matrix of fixed X values with first column of 1's and each of k vectors above comprising subsequent columns.
- **$X^{-1}$** is the Inverse Matrix of X, such that $X^{-1}X = I$ (the identity matrix).
- **$X^T$** is the transpose matrix of X where rows and columns are reversed.
- **Y** is the vector of $Y_i$'s arrayed as a column of numbers.
- **b** is the vector of regression coefficients including $\alpha$ plus all $\beta_i$'s arrayed as a single column of numbers.
- **$Y_{hat}$** is the vector of fitted values for $Y_i$ arrayed as a column of numbers.
- **e** is the vector of residuals $e_i$ ($Yhat_i$ - $Y_i$) arrayed as a column of numbers.

$$X = \begin{pmatrix} 1 & 6 & 9.9 & 5.7 & 1.6 \\ 1 & 1 & 9.3 & 6.4 & 3 \\ 1 & -2 & 9.4 & 5.7 & 3.4 \\ 1 & 11 & 9.1 & 6.1 & 3.4 \\ 1 & -1 & 6.9 & 6 & 3 \\ 1 & 2 & 9.3 & 5.7 & 4.4 \\ 1 & 5 & 7.9 & 5.9 & 2.2 \\ 1 & 1 & 7.4 & 6.2 & 2.2 \\ 1 & 1 & 7.3 & 5.5 & 1.9 \\ 1 & 3 & 8.8 & 5.2 & 0.2 \\ 1 & 11 & 9.8 & 5.7 & 4.2 \\ 1 & 9 & 10.5 & 6.1 & 2.4 \\ 1 & 5 & 9.1 & 6.4 & 3.4 \\ 1 & -3 & 10.1 & 5.5 & 3 \\ 1 & 1 & 7.2 & 5.5 & 0.2 \\ 1 & 8 & 11.7 & 6 & 3.9 \\ 1 & -2 & 8.7 & 5.5 & 2.2 \\ 1 & 3 & 7.6 & 6.2 & 4.4 \\ 1 & 6 & 8.6 & 5.9 & 0.2 \\ 1 & 10 & 10.9 & 5.6 & 2.4 \\ 1 & 4 & 7.6 & 5.8 & 2.4 \\ 1 & 5 & 7.3 & 5.8 & 4.4 \\ 1 & 5 & 9.2 & 5.2 & 1.6 \\ 1 & 3 & 7 & 6 & 1.9 \\ 1 & 8 & 7.2 & 5.5 & 1.6 \\ 1 & 8 & 7 & 6.4 & 4.1 \\ 1 & 6 & 8.8 & 6.2 & 1.9 \\ 1 & 6 & 10.1 & 5.4 & 2.2 \\ 1 & 3 & 12.1 & 5.4 & 4.1 \\ 1 & 5 & 7.7 & 6.2 & 1.6 \\ 1 & 1 & 7.8 & 6.8 & 2.4 \\ 1 & 8 & 11.5 & 6.2 & 1.9 \\ 1 & 10 & 10.4 & 6.4 & 2.2 \end{pmatrix}$$

## Constructing Design Matrix X of Independent Variables:

$X0_i := 1$ **< vector of 1's for constructing matrix X below**

$X := \text{augment}(X0, X1, X2, X3, X4)$

## Estimated Regression Coefficients (b):

$$b := \left(X^T X\right)^{-1} \cdot \left(X^T Y\right)$$

**^ Note: all calculations here involve MathCad's matrix algebra functions!**

$$b = \begin{pmatrix} 2.958283 \\ -0.129319 \\ -0.018785 \\ -0.046215 \\ 0.208755 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta 1 \\ \beta 2 \\ \beta 3 \\ \beta 4 \end{pmatrix}$$

## Estimated values of Y ($Y_{hat}$):

$Y_{hat} := X \cdot b$ **see next page for $Y_{hat}$**

$$X^T X = \begin{pmatrix} 33 & 147 & 293.2 & 194.1 & 83.9 \\ 147 & 1127 & 1366.1 & 872.7 & 381.3 \\ 293.2 & 1366.1 & 2675.66 & 1721.84 & 755.6 \\ 194.1 & 872.7 & 1721.84 & 1146.57 & 497.07 \\ 83.9 & 381.3 & 755.6 & 497.07 & 258.27 \end{pmatrix}$$

**^ Sums of Squares and Cross-Products of X's Matrix analog of Lxx**

## Residuals ($\varepsilon$):

$\varepsilon := Y - Y_{hat}$ **see next page for $\varepsilon$**

$$X^T Y = \begin{pmatrix} 81.65 \\ 302.73 \\ 718.605 \\ 479.779 \\ 215.64 \end{pmatrix}$$

**< Sums of XY cross Products Matrix of Lxy**

$$\left(X^T X\right)^{-1} = \begin{pmatrix} 10.351 & 0.0441 & -0.2354 & -1.4788 & 0.1072 \\ 0.0441 & 0.0025 & -0.0025 & -0.0059 & 0.0006 \\ -0.2354 & -0.0025 & 0.0176 & 0.0174 & -0.005 \\ -1.4788 & -0.0059 & 0.0174 & 0.2392 & -0.022 \\ 0.1072 & 0.0006 & -0.005 & -0.022 & 0.025 \end{pmatrix}$$

**< Inverse of Matrix $X^TX$: Multiplication by this matrix is the matrix algebra analog of dividing by $X^TX$.**

## Sums of Squares as Quadratic Forms:

$H := X \cdot \left(X^T \cdot X\right)^{-1} \cdot X^T$ **< called the "Hat Matrix" often reported in software**

$J_{i,ii} := 1$ **< nXn matrix of 1's useful in calculations**

$I := \text{identity}(n)$ **< identity matrix of length n**

$SSR := Y^T \cdot \left[H - \left(\frac{1}{n}\right) \cdot J\right] \cdot Y$ $SSR = (9.7174)$

$SSE := Y^T \cdot (I - H) \cdot Y$ $SSE = (5.0299)$

$SSTO := Y^T \cdot \left[I - \left(\frac{1}{n}\right) \cdot J\right] \cdot Y$ $SSTO = (14.747206)$

## Degrees of Freedom:

$df_R := k - 1$ $df_R = 4$

$df_E := n - k$ $df_E = 28$

$df_T := n - 1$ $df_T = 32$

## ANOVA Table:

| SS | df | MS | |
|---|---|---|---|
| $SSR = (9.7174)$ | $df_R = 4$ | $MSR := \frac{SSR}{df_R}$ | $MSR = (2.4293)$ |
| $SSE = (5.0299)$ | $df_E = 28$ | $MSE := \frac{SSE}{df_E}$ | $MSE = (0.1796)$ |
| $SSTO = (14.7472)$ | $df_T = 32$ | $MSTO := \frac{SSTO}{df_T}$ | $MSTO = (0.4609)$ |

| $H \cdot Y =$ | $Y_{hat} =$ | $\varepsilon =$ |
|---|---|---|
| 2.067 | 2.067 | 0.053 |
| 2.9848 | 2.9848 | 0.4052 |
| 3.4867 | 3.4867 | 0.1233 |
| 1.7927 | 1.7927 | −0.0727 |
| 3.307 | 3.307 | −1.507 |
| 3.18 | 3.18 | 0.03 |
| 2.3499 | 2.3499 | 0.2401 |
| 2.8627 | 2.8627 | 0.3873 |
| 2.8343 | 2.8343 | 0.0257 |
| 2.2065 | 2.2065 | 0.1135 |
| 1.965 | 1.965 | −0.395 |
| 1.8163 | 1.8163 | −0.3163 |
| 2.5547 | 2.5547 | 0.1353 |
| 3.5286 | 3.5286 | 0.5314 |
| 2.4813 | 2.4813 | −0.5013 |
| 2.2408 | 2.2408 | 0.0492 |
| 3.2586 | 3.2586 | 0.2914 |
| 3.0596 | 3.0596 | 0.2504 |
| 1.7899 | 1.7899 | 0.0401 |
| 1.7025 | 1.7025 | −0.0125 |
| 2.5312 | 2.5312 | −0.1112 |
| 2.825 | 2.825 | 0.155 |
| 2.2326 | 2.2326 | −0.3926 |
| 2.5582 | 2.5582 | −0.0782 |
| 1.8683 | 1.8683 | 0.9617 |
| 2.3524 | 2.3524 | 0.0576 |
| 2.1272 | 2.1272 | −0.3472 |
| 2.2023 | 2.2023 | 0.0177 |
| 2.9494 | 2.9494 | −0.2294 |
| 2.2145 | 2.2145 | 0.1455 |
| 2.8692 | 2.8692 | −0.0592 |
| 1.8178 | 1.8178 | −0.1778 |
| 1.6332 | 1.6332 | 0.1868 |

## Standard error of the Regression Parameters:

$MS_E := MSE_0$ $MS_E = 0.1796$ **< converting 1X1 matrix into scalar**

$sb_{sq} := MS_E \cdot \left(X^T \cdot X\right)^{-1}$ **< matrix of variances/covariances among regression coefficients:**

$sb_j := \sqrt{sb_{sq_{j,j}}}$

$$sb = \begin{pmatrix} 1.363608 \\ 0.021287 \\ 0.056278 \\ 0.20727 \\ 0.067035 \end{pmatrix}$$

**< standard error of regression parameters b**

$$sb_{sq} = \begin{pmatrix} 1.8594 & 0.0079 & -0.0423 & -0.2656 & 0.0193 \\ 0.0079 & 0.0005 & -0.0004 & -0.0011 & 0.0001 \\ -0.0423 & -0.0004 & 0.0032 & 0.0031 & -0.0009 \\ -0.2656 & -0.0011 & 0.0031 & 0.043 & -0.004 \\ 0.0193 & 0.0001 & -0.0009 & -0.004 & 0.0045 \end{pmatrix}$$

## Coefficient of Multiple Determination ($R^2$):

$R_{sq} := 1 - \frac{SSE_0}{SSTO_0}$ $R_{sq} = 0.65893$

## Coefficient of MultipleCorrelation (R):

$R := \sqrt{R_{sq}}$ $R = 0.8117$

## Adjusted Coefficient of Multiple Determination:

$R_{sqa} := 1 - \frac{MSE_0}{MSTO_0}$ $R_{sqa} = 0.6102$

## Overall F Test of the Multiple Regression:

### Hypotheses:

**$H_0$: all slope $\beta$'s = 0** **< i.e., only $\alpha$ is left in regression model**

**$H_1$: at least some slope $\beta$'s not zero**

### Test Statistic:

$F := \frac{MSR_0}{MSE_0}$ $F = 13.5235$

### Sampling Distribution:

**If Assumptions hold and $H_0$ is true, then F ~$F_{(k-1)/(n-k)}$**

### Critical Value of the Test:

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

$C := qF(1 - \alpha, k - 1, n - k)$ $C = 2.7141$

### Decision Rule:

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

$F = 13.5235$ $C = 2.7141$

### Probability Value:

$P := 1 - pF(F, k - 1, n - k)$ $P = 2.94784 \times 10^{-6}$

## Partial t/F-Tests of single coefficients:

**Note: this is a "marginal" test, so the order of entry into regression does not matter.**

### Hypotheses:

**$H_0$: a single $\beta = 0$** **< typically this is the marginal independent variable, but also intercept $\alpha$**

**$H_1$: $\beta_j \neq 0$** **< test set each taking a turn ($\alpha$ first, then all $\beta$'s in turn)**

### Test Statistic:

$$t_j := \frac{b_j}{sb_j}$$

$$F_j := \left(t_j\right)^2$$

$$t = \begin{pmatrix} 2.1695 \\ -6.0751 \\ -0.3338 \\ -0.223 \\ 3.1141 \end{pmatrix} \quad F = \begin{pmatrix} 4.7065 \\ 36.9063 \\ 0.1114 \\ 0.0497 \\ 9.6979 \end{pmatrix} \quad t^2 = \begin{pmatrix} 4.7065 \\ 36.9063 \\ 0.1114 \\ 0.0497 \\ 9.6979 \end{pmatrix}$$

### Sampling Distributions:

**If Assumptions hold and $H_0$ is true, then $t \sim t_{(n-k)}$**

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

$k = 5$

### Critical Values of the Test:

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

$$CV_t := qt\left(1 - \frac{\alpha}{2}, n - k\right) \qquad CV_t = 2.0484$$

$$CV_F := qF\left(1 - \frac{\alpha}{2}, 1, n - k\right) \qquad CV_F = 5.6096$$

### Decision Rules:

**IF $|t| > CV_t$, THEN REJECT $H_0$ OTHERWISE ACCEPT $H_0$** $CV_t = 2.0484$ **< t-test version**

**IF $F > CV_F$, THEN REJECT $H_0$ OTHERWISE ACCEPT $H_0$** $CV_F = 5.6096$ **< F-test version**

### Probability Values:

$$P_{t_j} := \min\left[2 \cdot pt\left(t_j, n - k\right), 2 \cdot \left(1 - pt\left(t_j, n - k\right)\right)\right]$$

**^ takes the minimum of the two-tailed P calculations**

$$P_{F_j} := 1 - pF\left(F_j, 1, n - k\right)$$

$$P_t = \begin{pmatrix} 0.038691 \\ 0.000001 \\ 0.741024 \\ 0.825178 \\ 0.004227 \end{pmatrix} \quad \textbf{for: } \alpha,\ \beta1,\ \beta2,\ \beta3,\ \beta4 \quad P_F = \begin{pmatrix} 0.038691 \\ 0.000001 \\ 0.741024 \\ 0.825178 \\ 0.004227 \end{pmatrix}$$

## Confidence Intervals for single coefficients $\alpha$ & $\beta$:

$CI_b := \text{augment}(b - C \cdot sb, b + C \cdot sb)$ **< augment function puts them into a matrix for display**

$$CI_b = \begin{pmatrix} -0.7427 & 6.6592 \\ -0.1871 & -0.0715 \\ -0.1715 & 0.134 \\ -0.6088 & 0.5163 \\ 0.0268 & 0.3907 \end{pmatrix} \qquad b = \begin{pmatrix} 2.9583 \\ -0.1293 \\ -0.0188 \\ -0.0462 \\ 0.2088 \end{pmatrix}$$

**for:**
**$\alpha$**
**$\beta 1$**
**$\beta 2$**
**$\beta 3$**
**$\beta 4$**

## Prototype in R:

```
#MULTIPLE REGRESSION
#ZAR EXAMPLE 20.1a
ZAR=read.table("ZarEX20.1aR.txt")
ZAR
attach(ZAR)

Y=var5
X1=var1
X2=var2
X3=var3
X4=var4

LM=lm(Y~X1+X2+X3+X4)
LM

summary(LM)
anova(LM)
```

```
> LM

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4)

Coefficients:
(Intercept)           X1           X2           X3           X4
    2.95828     -0.12932     -0.01878     -0.04621      0.20876
```

```
> summary(LM)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5070 -0.1112  0.0401  0.1550  0.9617

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.95828    1.36361   2.169  0.03869 *
X1          -0.12932    0.02129  -6.075 1.50e-06 ***
X2          -0.01878    0.05628  -0.334  0.74102
X3          -0.04621    0.20727  -0.223  0.82518
X4           0.20876    0.06703   3.114  0.00423 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4238 on 28 degrees of freedom
Multiple R-squared: 0.6589,     Adjusted R-squared: 0.6102
F-statistic: 13.52 on 4 and 28 DF,  p-value: 2.948e-06
```

**Partial t-tests of individual coefficients >**

**Overall F-test >**

**Coefficient of Multiple Determination $R^2$ ^**

**Adjusted Coefficient of Multiple Determination $R_a^2$ ^**

$$b = \begin{pmatrix} 2.9583 \\ -0.1293 \\ -0.0188 \\ -0.0462 \\ 0.2088 \end{pmatrix} \quad sb = \begin{pmatrix} 1.36361 \\ 0.02129 \\ 0.05628 \\ 0.20727 \\ 0.06703 \end{pmatrix} \quad t = \begin{pmatrix} 2.1695 \\ -6.0751 \\ -0.3338 \\ -0.223 \\ 3.1141 \end{pmatrix} \quad P_t = \begin{pmatrix} 0.0387 \\ 1.4958 \times 10^{-6} \\ 0.741 \\ 0.8252 \\ 0.0042 \end{pmatrix}$$

**< equivalent to summary() report of partial t-tests of individual coefficients**

**Note: Sums of Squares for X1-X4 in R's ANOVA table sum to SSR above, and SS for 'Residuals' in R equals SSE above.**

```
> anova(LM)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 7.8762  7.8762 43.8450 3.496e-07 ***
X2         1 0.0131  0.0131  0.0732  0.788785
X3         1 0.0859  0.0859  0.4781  0.494963
X4         1 1.7421  1.7421  9.6979  0.004227 **
Residuals 28 5.0299  0.1796
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Sums of Squares for X1-X4 in the ANOVA chart, also known as "Partial" or "Extra" Sums of Squares are not calculated in this worksheet. To understand how to interpret them, see *Biostatistics* Worksheet 400.**