ORIGIN ≡ 0

Multiple Linear Regression

W. Stein

Linear regression models are easily extended to include multiple vectors of independent variables. If the response (dependent) variable remains a single vector, then the models are called "multiple regression". If, in addition to multiple vectors of independent variables, the response consists of multiple vectors of dependent variables, the models are called "multivariate multiple regresssion" Shown here is multiple regression in matrix form. Terminology is drawn from Kuter et al. (KNNL) A*pplied Linear Statistical Models* 5th Edition. The example comes from Peter Dalgaard (PD) *Introductory Statistics with R* and his ISwR package available for download at any R CRAN site.

Assumptions:

- Multiple 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.

- $X_{1,0}, X_{2,0}, X_{3,0}, \ldots, X_{n,0}$ (first independent variable) with each value of X_{i0} matched to Y_{i0}

- $X_{1,1}, X_{2,1}, X_{3,1}, \ldots, X_{n,1}$ (second independent variable) with each value of X_{i1} matched to Y_{i1}

...

- $X_{1,j}, X_{2,j}, X_{3,j}, \ldots, X_{n,j}$ (last - j^{th} independent variable = (p-1) with each value of $X_{i,j}$ matched to $Y_{i,j}$

Model:

$$Y_i = \beta_0 + \Sigma\beta_j X_{n,j} + \varepsilon_i$$

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

β_j is the slope of the regression line (scaling coefficient) for each $X_{,j}$,

ε_i is the error factor in prediction of Y_i and

a random variable distributed as $\sigma^2 I$ - for I=the Identity matrix

Example:

Data comes from ISwR package as dataset "cystfibr". This example is worked extensively in Dalgaard and is large enough (with several independent variables) to be an interesting worked example showing how matrix algebra works with real data. To obtain the dataset in R, one must download the ISwR package from a CRAN site. Then from the R script below will load the data, and assign variables. All calculations in MathCad are based on data matrix K read from a file I've constructed from ISwR. To use other data, modify the section below for K. Be sure to specify X & Y columns as done in Variable Assignment. After that, calculations should flow properly.

K := READPRN("c:/2008LinearModelsData/cystfibrM.txt")

Variable Assignment: **using original variable names**

$pemax := K^{\langle 9\rangle}$ **< dependent variable assignment**

$age := K^{\langle 0\rangle}$	$weight := K^{\langle 3\rangle}$	$rv := K^{\langle 6\rangle}$
$sex := K^{\langle 1\rangle}$	$bmp := K^{\langle 4\rangle}$	$frc := K^{\langle 7\rangle}$
$height := K^{\langle 2\rangle}$	$fev1 := K^{\langle 5\rangle}$	$tlc := K^{\langle 8\rangle}$

< independent variable assignments

Summary Statistics:

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

p := 10 **< number of independent variables - must be explicitly set.**

Range Variables:

$i := 0 .. n - 1$

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

$j := 0 .. p - 1$ **< range variable j for columns of X**

Matrix Formulation:

$Y := pemax$ **< dependent vector**

$OV_i := 1$ **< vector of 1**

$X := augment(OV, age, sex, height, weight, bmp, fev1, rv, frc, tlc)$

^ augment function puts the design matrix together

$I := identity(n)$ **< identity matrix of length n**

$J_{i,ii} := 1$ **< square matrix of 1 of size n X n**

X =

1	7	0	109	13.1	68	32	258	183	137
1	7	1	112	12.9	65	19	449	245	134
1	8	0	124	14.1	64	22	441	268	147
1	8	1	125	16.2	67	41	234	146	124
1	8	0	127	21.5	93	52	202	131	104
1	9	0	130	17.5	68	44	308	155	118
1	11	1	139	30.7	89	28	305	179	119
1	12	1	150	28.4	69	18	369	198	103
1	12	0	146	25.1	67	24	312	194	128
1	13	1	155	31.5	68	23	413	225	136
1	13	0	156	39.9	89	39	206	142	95
1	14	1	153	42.1	90	26	253	191	121
1	14	0	160	45.6	93	45	174	139	108
1	15	1	158	51.2	93	45	158	124	90
1	16	1	160	35.9	66	31	302	133	101
1	17	1	153	34.8	70	29	204	118	120
1	17	0	174	44.7	70	49	187	104	103
1	17	1	176	60.1	92	29	188	129	130
1	17	0	171	42.6	69	38	172	130	103
1	19	1	156	37.2	72	21	216	119	81
1	19	0	174	54.6	86	37	184	118	101
1	20	0	178	64	86	34	225	148	135
1	23	0	180	73.8	97	57	171	108	98
1	23	0	175	51.1	71	33	224	131	113
1	23	0	179	71.5	95	52	225	127	101

Y =

95
85
100
85
95
80
65
110
70
95
110
90
100
80
134
134
165
120
130
85
85
160
165
95
195

K =

7	0	109	13.1	68	32	258	183	137	95
7	1	112	12.9	65	19	449	245	134	85
8	0	124	14.1	64	22	441	268	147	100
8	1	125	16.2	67	41	234	146	124	85
8	0	127	21.5	93	52	202	131	104	95
9	0	130	17.5	68	44	308	155	118	80
11	1	139	30.7	89	28	305	179	119	65
12	1	150	28.4	69	18	369	198	103	110
12	0	146	25.1	67	24	312	194	128	70
13	1	155	31.5	68	23	413	225	136	95
13	0	156	39.9	89	39	206	142	95	110
14	1	153	42.1	90	26	253	191	121	90
14	0	160	45.6	93	45	174	139	108	100
15	1	158	51.2	93	45	158	124	90	80
16	1	160	35.9	66	31	302	133	101	134
17	1	153	34.8	70	29	204	118	120	134
17	0	174	44.7	70	49	187	104	103	165
17	1	176	60.1	92	29	188	129	130	120
17	0	171	42.6	69	38	172	130	103	130
19	1	156	37.2	72	21	216	119	81	85
19	0	174	54.6	86	37	184	118	101	85
20	0	178	64	86	34	225	148	135	160
23	0	180	73.8	97	57	171	108	98	165
23	0	175	51.1	71	33	224	131	113	95
23	0	179	71.5	95	52	225	127	101	195

Least Squares Estimation of the Regression Parameters:

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

^ Note *error* in KNNL Eq. 6.25
It should read:
$b=(X'X)^{-1}X'Y$

b =

176.0582
–2.542
–3.7368
–0.4463
2.9928
–1.7449
1.0807
0.197
–0.3084
0.1886

$X^T \cdot Y =$

2728
41992
1083
427177
113889.8
215759
98790
674404
409294
308519

^ values verified PD p. 151-152

$$\left(X^T \cdot X\right)^{-1} = \begin{pmatrix} 78.6541 & -1.0458 & -2.2695 & -0.2508 & 0.6214 & -0.2927 & -0.2212 & 0.0039 & -0.0501 & -0.0934 \\ -1.0458 & 0.0355 & 0.0421 & 0.0011 & -0.0118 & 0.005 & 0.0039 & -0.0004 & 0.0015 & 0.0012 \\ -2.2695 & 0.0421 & 0.3684 & 0.0051 & -0.0156 & 0.0003 & 0.0182 & -0.0019 & 0.0059 & 0.002 \\ -0.2508 & 0.0011 & 0.0051 & 0.0013 & -0.0019 & 0.0007 & 0.0005 & -0 & 0.0001 & 0.0002 \\ 0.6214 & -0.0118 & -0.0156 & -0.0019 & 0.0062 & -0.0028 & -0.0016 & 0.0001 & -0.0004 & -0.0007 \\ -0.2927 & 0.005 & 0.0003 & 0.0007 & -0.0028 & 0.0021 & 0.0001 & 0.0001 & -0.0001 & 0.0004 \\ -0.2212 & 0.0039 & 0.0182 & 0.0005 & -0.0016 & 0.0001 & 0.0018 & -0.0001 & 0.0005 & 0.0001 \\ 0.0039 & -0.0004 & -0.0019 & -0 & 0.0001 & 0.0001 & -0.0001 & 0.0001 & -0.0001 & 0 \\ -0.0501 & 0.0015 & 0.0059 & 0.0001 & -0.0004 & -0.0001 & 0.0005 & -0.0001 & 0.0004 & -0.0001 \\ -0.0934 & 0.0012 & 0.002 & 0.0002 & -0.0007 & 0.0004 & 0.0001 & 0 & -0.0001 & 0.0004 \end{pmatrix}$$

Fitted Values & Hat Matrix H:

$Y_h := X \cdot b$ **< using standard way to calculate Y_h**

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

^ using H to find Y_h

H =

	0	1	2	3	4
0	0.5773	0.2222	0.1353	0.1718	0.016
1	0.2222	0.4334	0.1898	0.0755	-0.0856
2	0.1353	0.1898	0.423	-0.0297	-0.006
3	0.1718	0.0755	-0.0297	0.4586	0.0663
4	0.016	-0.0856	-0.006	0.0663	0.558
5	0.0889	0.0928	0.0496	0.1086	0.2286
6	-0.0749	0.0679	0.014	-0.0139	0.247
7	-0.0873	0.1827	0.0588	-0.084	-0.1011
8	0.0978	0.0061	0.1937	-0.0642	0.0164
9	-0.2469	0.0715	0.2078	0.075	-0.0509
10	0.0179	-0.039	0.0109	-0.1205	0.1697
11	-0.0391	-0.0071	0.1202	0.0539	0.0301

^ hat matrix n X n

$$Y_h = \begin{pmatrix} 84.969 \\ 88.4138 \\ 86.6141 \\ 96.5319 \\ 76.3086 \\ 111.5517 \\ 76.4802 \\ 89.9663 \\ 90.3072 \\ 108.1822 \\ 94.3536 \\ 79.252 \\ 103.6641 \\ 113.1177 \\ 123.5396 \\ 100.5948 \\ 143.9664 \\ 123.0021 \\ 117.9037 \\ 83.9192 \\ 122.3377 \\ 148.1364 \\ 169.3318 \\ 129.2326 \\ 166.3231 \end{pmatrix} \quad Y_{hh} = \begin{pmatrix} 84.969 \\ 88.4138 \\ 86.6141 \\ 96.5319 \\ 76.3086 \\ 111.5517 \\ 76.4802 \\ 89.9663 \\ 90.3072 \\ 108.1822 \\ 94.3536 \\ 79.252 \\ 103.6641 \\ 113.1177 \\ 123.5396 \\ 100.5948 \\ 143.9664 \\ 123.0021 \\ 117.9037 \\ 83.9192 \\ 122.3377 \\ 148.1364 \\ 169.3318 \\ 129.2326 \\ 166.3231 \end{pmatrix}$$

^ Y_h calculated either way...

Residuals:

$e := Y - Y_h$ **< standard definition of residual**

$ee := (I - H) \cdot Y$ **< Eq. 6.31 KNNL**

^ see list of residuals on next page...

Sums of Squares as Quadratic Forms:

$SSR := Y^T \cdot \left[H - \left(\frac{1}{n}\right) \cdot J\right] \cdot Y$ $\quad SSR = (17101.3904)$

$SSE := Y^T \cdot (I - H) \cdot Y$ $\quad SSE = (9731.2496)$

$SSTO := Y^T \cdot \left[I - \left(\frac{1}{n}\right) \cdot J\right] \cdot Y$ $\quad SSTO = (26832.64)$

Degrees of Freedom:

$df_R := p - 1$ $\quad df_R = 9$

$df_E := n - p$ $\quad df_E = 15$

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

ANOVA Table:

SS	df	MS	
$SSR = (17101.3904)$	$df_R = 9$	$MSR := \frac{SSR}{df_R}$	$MSR = (1900.1545)$
$SSE = (9731.2496)$	$df_E = 15$	$MSE := \frac{SSE}{df_E}$	$MSE = (648.75)$
$SSTO = (26832.64)$	$df_T = 24$	$MSTO := \frac{SSTO}{df_T}$	$MSTO = (1118.0267)$

^ verified RD p. 152 (by adding partial SSR's)

$$e = \begin{pmatrix} 10.031 \\ -3.4138 \\ 13.3859 \\ -11.5319 \\ 18.6914 \\ -31.5517 \\ -11.4802 \\ 20.0337 \\ -20.3072 \\ -13.1822 \\ 15.6464 \\ 10.748 \\ -3.6641 \\ -33.1177 \\ 10.4604 \\ 33.4052 \\ 21.0336 \\ -3.0021 \\ 12.0963 \\ 1.0808 \\ -37.3377 \\ 11.8636 \\ -4.3318 \\ -34.2326 \\ 28.6769 \end{pmatrix} \quad ee = \begin{pmatrix} 10.031 \\ -3.4138 \\ 13.3859 \\ -11.5319 \\ 18.6914 \\ -31.5517 \\ -11.4802 \\ 20.0337 \\ -20.3072 \\ -13.1822 \\ 15.6464 \\ 10.748 \\ -3.6641 \\ -33.1177 \\ 10.4604 \\ 33.4052 \\ 21.0336 \\ -3.0021 \\ 12.0963 \\ 1.0808 \\ -37.3377 \\ 11.8636 \\ -4.3318 \\ -34.2326 \\ 28.6769 \end{pmatrix}$$

^ residuals calculated either way

Variance/Covariance Matrix of Residuals:

$s_{sq} := MSE_0 \cdot (I - H)$ **< MSE_0 is MathCad's way of converting from a 1 X 1 matrix to a scalar**

see KNNL Eq. 6.33

$s_{sq} =$

	0	1	2	3	4	5	6
0	274.2184	-144.1633	-87.7897	-111.4471	-10.3798	-57.6612	48.6016
1	-144.1633	367.5818	-123.1125	-48.9731	55.5122	-60.1754	-44.04
2	-87.7897	-123.1125	374.3183	19.2908	3.8804	-32.1714	-9.1131
3	-111.4471	-48.9731	19.2908	351.2163	-42.9982	-70.4582	9.0278
4	-10.3798	55.5122	3.8804	-42.9982	286.7768	-148.2891	-160.2523
5	-57.6612	-60.1754	-32.1714	-70.4582	-148.2891	427.3583	-26.0804
6	48.6016	-44.04	-9.1131	9.0278	-160.2523	-26.0804	448.8841
7	56.6377	-118.5534	-38.1483	54.5077	65.6116	-14.202	-32.0713
8	-63.4788	-3.9581	-125.6827	41.6321	-10.6199	-32.3342	-2.6856
9	160.2036	-46.4033	-134.8266	-48.6819	33.0108	-13.3829	-73.0739

^ n X n matrix of variances/covariances

Variance/Covariance Matrix of Coefficients:

$$sb_{sq} := MSE_0 \cdot \left(X^T \cdot X\right)^{-1}$$

$$sb_{sq} = \begin{pmatrix} 51026.8157 & -678.4945 & -1472.3482 & -162.7221 & 403.1311 & -189.8869 & -143.5257 & 2.5595 & -32.4731 & -60.5758 \\ -678.4945 & 23.0563 & 27.3396 & 0.6819 & -7.6471 & 3.2597 & 2.5623 & -0.2284 & 0.9446 & 0.7847 \\ -1472.3482 & 27.3396 & 239.0061 & 3.2888 & -10.0998 & 0.2115 & 11.7939 & -1.2196 & 3.8446 & 1.2666 \\ -162.7221 & 0.6819 & 3.2888 & 0.8161 & -1.2245 & 0.4742 & 0.3403 & -0.0106 & 0.0831 & 0.1279 \\ 403.1311 & -7.6471 & -10.0998 & -1.2245 & 4.0319 & -1.821 & -1.0274 & 0.055 & -0.2753 & -0.4276 \\ -189.8869 & 3.2597 & 0.2115 & 0.4742 & -1.821 & 1.3346 & 0.0844 & 0.0574 & -0.0739 & 0.2587 \\ -143.5257 & 2.5623 & 11.7939 & 0.3403 & -1.0274 & 0.0844 & 1.1684 & -0.0714 & 0.3239 & 0.0826 \\ 2.5595 & -0.2284 & -1.2196 & -0.0106 & 0.055 & 0.0574 & -0.0714 & 0.0385 & -0.082 & 0.0149 \\ -32.4731 & 0.9446 & 3.8446 & 0.0831 & -0.2753 & -0.0739 & 0.3239 & -0.082 & 0.2424 & -0.0634 \\ -60.5758 & 0.7847 & 1.2666 & 0.1279 & -0.4276 & 0.2587 & 0.0826 & 0.0149 & -0.0634 & 0.2497 \end{pmatrix}$$

^ p X p matrix of variances/covariances among the b's

Vector of Standard deviations:

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

$$sb = \begin{pmatrix} 225.8912 \\ 4.8017 \\ 15.4598 \\ 0.9034 \\ 2.008 \\ 1.1552 \\ 1.0809 \\ 0.1962 \\ 0.4924 \\ 0.4997 \end{pmatrix}$$

S

< standard deviation of Y's for each b

This is the square-root on the main diagonal of sb_{sq}

$\sqrt{s}$

Coefficient of Multiple Determination (R^2):

$$R_{sq} := 1 - \frac{SSE_0}{SSTO_0} \qquad R_{sq} = 0.63734$$

^ verified RD p. 152

Coefficient of MultipleCorrelation (R):

$$R := \sqrt{R_{sq}} \qquad R = 0.7983$$

Adjusted Coefficient of Multiple Determination:

$$R_{sqa} := 1 - \frac{MSE_0}{MSTO_0} \qquad R_{sqa} = 0.4197$$

< confirmed PD p. 152

$$1 - \left(\frac{n-1}{n-p}\right) \cdot \left(\frac{SSE_0}{SSTO_0}\right) = 0.4197$$

< alternate formula gives same answer here

Overall F Test of Regression:

Critical Value of the Test:

Hypotheses:

H_0: all slope β's = 0 **< i.e., only β_0 left in model**

H_1: at least some slope β's not zero

Test Statistic:

$F := \frac{MSR_0}{MSE_0}$ $F = 2.9289$ **< confirmed PD p. 152**

Critical Value of the Test:

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

$CV := qF(1 - \alpha, p - 1, n - p)$ $CV = 2.5876$

Decision Rule:

IF F > CV, THEN REJECT H_0 OTHERWISE ACCEPT H_0

$F = 2.9289$ $CV = 2.5876$

Probability Value:

$P := 1 - pF(F, p - 1, n - p)$ $P = 0.03195$ **< confirmed PD p. 152**

Partial t/F Test 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**

H_1: β_j <> 0

$k := j$ **< test set each taking a turn**

Test Statistic:

$$t_k := \frac{b_k}{sb_k}$$

$$F := t^2$$

$$t = \begin{pmatrix} 0.7794 \\ -0.5294 \\ -0.2417 \\ -0.494 \\ 1.4905 \\ -1.5105 \\ 0.9998 \\ 1.0039 \\ -0.6264 \\ 0.3774 \end{pmatrix}$$

< t statistics confirmed PD p. 152

$$F = \begin{pmatrix} 0.6075 \\ 0.2803 \\ 0.0584 \\ 0.244 \\ 2.2215 \\ 2.2815 \\ 0.9995 \\ 1.0077 \\ 0.3924 \\ 0.1424 \end{pmatrix}$$

< alternate F test statistics

Critical Value of the Test:

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

$CV_t := qt\left(1 - \frac{\alpha}{2}, n - p\right)$ $CV_t = 2.1314$

$CV_F := qF(1 - \alpha, 1, n - p)$ $CV_F = 4.5431$

Decision Rule:

IF |t| > CV_t, THEN REJECT H_0 OTHERWISE ACCEPT H_0 $CV_t = 2.1314$

IF F > CV_F, THEN REJECT H_0 OTHERWISE ACCEPT H_0 $CV_F = 4.5431$

Probability Value:

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

$P_{F_k} := 1 - pF(F_k, 1, n - p)$

$$P_t = \begin{pmatrix} 0.44787 \\ 0.60428 \\ 0.81228 \\ 0.62846 \\ 0.15683 \\ 0.1517 \\ 0.33328 \\ 0.33136 \\ 0.54047 \\ 0.71116 \end{pmatrix} \quad P_F = \begin{pmatrix} 0.44787 \\ 0.60428 \\ 0.81228 \\ 0.62846 \\ 0.15683 \\ 0.1517 \\ 0.33328 \\ 0.33136 \\ 0.54047 \\ 0.71116 \end{pmatrix}$$

< confirmed PD p. 152

Confidence Interval for single coefficients β:

$CI_b := augment(b - CV \cdot sb, b + CV \cdot sb)$

$$b = \begin{pmatrix} 176.0582 \\ -2.542 \\ -3.7368 \\ -0.4463 \\ 2.9928 \\ -1.7449 \\ 1.0807 \\ 0.197 \\ -0.3084 \\ 0.1886 \end{pmatrix} \quad CI_b = \begin{pmatrix} -408.4637 & 760.5801 \\ -14.967 & 9.883 \\ -43.741 & 36.2675 \\ -2.7838 & 1.8913 \\ -2.203 & 8.1887 \\ -4.7343 & 1.2444 \\ -1.7164 & 3.8778 \\ -0.3108 & 0.7047 \\ -1.5826 & 0.9657 \\ -1.1045 & 1.4817 \end{pmatrix}$$

GLM Test Approach for any subset of β's:

Full Model:

$Y_i = \beta_0 + \Sigma\beta_j X_{n,j} + \varepsilon_i$

Reduced Model:

$Y_i = \beta_0 + \beta_1 age + \beta_2 sex + \beta_4 weight + \beta_5 bmp + \varepsilon_i$ **< only some independent variables chosen for a test | change as desired**

$pR := 5$ **< note count set of parameters of the Reduced Model above**

Hypotheses:

H_0: Reduced Model that some but not all β's = 0

H_1: Full Model is required by the data

Error Sums of Squares for Full Model:

$SSE_F := SSE_0$ $SSE_F = 9731.2496$ **< from calculations above, and converting to a scalar.**

$df_F := df_E$ $df_F = 15$

Error Sums of Squares for Reduced Model:

$X_R :=$ augment(OV, age, sex, weight, bmp) **< making a design matrix for the Reduced Model**

$H_R := X_R \cdot \left(X_R^T \cdot X_R\right)^{-1} \cdot X_R^T$ **< calculating a Hat Matrix for the Reduced Model**

$SSE_R := \left[Y^T \cdot (I - H_R) \cdot Y\right]_0$ $SSE_R = 12602.1105$ **< calculating SSE for the Reduced Model and converting to a scalar**

$df_R := pR$ $df_R = 5$

Test Statistic:

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

$F = 0.885044$ **< verified using R: anova(FM,RM) - see script**

Critical Value of the Test:

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

$CV := qF(1 - \alpha, pR, n - p)$ $CV = 2.9013$

Decision Rule:

IF F > CV, THEN REJECT H_0 OTHERWISE ACCEPT H_0

$F = 0.885$ $CV = 2.9013$

Probability Value:

$P := 1 - pF(F, pR, n - p)$ $P = 0.5149$ **< verified using R: anova(FM,RM)**

Confidence/Prediction Regions for Regression Surface:

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

$X_h := X$ **< here using all original values of X & Y_h but any X values may be specified instead...**

Note that KNNL strangely define their "X_h" as the transpose of X_h here, see Eq. 6.53. Why they do this is not at all clear and very confusing. Thus, my formulas below differ from theirs in the use of transpose X_h, but is equivalent.

Critical Values:

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

$$CV_t := qt\left(1 - \frac{\alpha}{2}, n - p\right) \qquad CV_t = 2.1314$$

^ Note degrees of freedom = (n-p)

$$W := \sqrt{p \cdot qF(1 - \alpha, p, n - p)} \qquad W = 5.0435$$

Single Confidence Intervals CI_{Yh} for each X_n:

$$ss_{Y_i} := \left(X_h \cdot sb_{sq} \cdot X_h^T\right)_{i,i}$$ **< estimated standard deviation KNNL Eq. 6.57a**

$$CI_{Yh} := augment\left(Y_h - CV_t \cdot \sqrt{ss_Y}, Y_h + CV_t \cdot \sqrt{ss_Y}\right)$$ **< confidence Interval for Y_h given X_h**

^ values listed next page and verified by R:predict(FM,interval="confidence",level=0.95)

$$ss_Y = \begin{pmatrix} 374.5316 \\ 281.1681 \\ 274.4316 \\ 297.5337 \\ 361.9732 \\ 221.3917 \\ 199.8658 \\ 244.0893 \\ 138.4567 \\ 292.4531 \\ 184.9139 \\ 266.2465 \\ 135.1897 \\ 294.8515 \\ 221.4221 \\ 265.0592 \\ 234.2773 \\ 376.7031 \\ 259.6346 \\ 370.4377 \\ 124.9472 \\ 254.7846 \\ 268.224 \\ 273.8648 \\ 271.0487 \end{pmatrix}$$

^ see KNNL Eq. 6.59

$$MSE_0 \cdot \left[X_h \cdot \left(X^T \cdot X\right)^{-1} \cdot X_h^T\right] =$$

	0	1	2	3
0	374.5316	144.1633	87.7897	111.4471
1	144.1633	281.1681	123.1125	48.9731
2	87.7897	123.1125	274.4316	-19.2908
3	111.4471	48.9731	-19.2908	297.5337
4	10.3798	-55.5122	-3.8804	42.9982
5	57.6612	60.1754	32.1714	70.4582
6	-48.6016	44.04	9.1131	-9.0278

$$X_h \cdot sb_{sq} \cdot X_h^T =$$

	0	1	2	3
0	374.5316	144.1633	87.7897	111.4471
1	144.1633	281.1681	123.1125	48.9731
2	87.7897	123.1125	274.4316	-19.2908
3	111.4471	48.9731	-19.2908	297.5337
4	10.3798	-55.5122	-3.8804	42.9982
5	57.6612	60.1754	32.1714	70.4582
6	-48.6016	44.04	9.1131	-9.0278

^ Equivalent - See KNNL Eq. 6.58

Single Prediction Intervals PI_{Yh} for each X_n:

$$PI_{Yh} := \text{augment}\left(Y_h - CV_t \cdot \sqrt{MSE_0 + ss_Y}, Y_h + CV_t \cdot \sqrt{MSE_0 + ss_Y}\right)$$

Working-Hotelling Simultaneous Confidence Band WI_{Yh} for all X_n:

$$WI_{Yh} := \text{augment}\left(Y_h - \sqrt{W} \cdot \sqrt{ss_Y}, Y_h + \sqrt{W} \cdot \sqrt{ss_Y}\right)$$

^ values not confirmed yet, but seem reasonable...

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

$Y_h =$	$CI_{Yh} =$		$PI_{Yh} =$		$WI_{Yh} =$	
84.969	43.71943	126.21854	16.7865	153.1514	41.5068	128.4312
88.4138	52.67346	124.15405	23.4161	153.4114	50.7564	126.0712
86.6141	51.30451	121.9236	21.8523	151.3758	49.4105	123.8176
96.5319	59.7662	133.29764	30.9649	162.099	57.7941	135.2698
76.3086	35.75649	116.86067	8.5458	144.0713	33.5813	119.0359
111.5517	79.83738	143.26606	48.6779	174.4256	78.1362	144.9672
76.4802	46.34711	106.61338	14.389	138.5715	44.7308	108.2297
89.9663	56.6659	123.26669	26.2777	153.6549	54.8797	125.0529
90.3072	65.22697	115.3875	30.5047	150.1098	63.8817	116.7328
108.1822	71.73172	144.63267	42.7914	173.573	69.7765	146.5879
94.3536	65.36953	123.33773	32.8118	155.8955	63.8148	124.8924
79.252	44.47304	114.03102	14.778	143.7261	42.6075	115.8966
103.6641	78.88152	128.44671	43.9858	163.3424	77.5522	129.776
113.1177	76.51811	149.71738	47.6437	178.5918	74.5549	151.6806
123.5396	91.82311	155.25615	60.6647	186.4146	90.1218	156.9574
100.5948	65.89347	135.29619	36.1626	165.027	64.0321	137.1576
143.9664	111.34213	176.59058	80.6287	207.304	109.5922	178.3405
123.0021	81.63318	164.37111	54.7474	191.2569	79.4142	166.5901
117.9037	83.55925	152.24812	53.663	182.1443	81.717	154.0904
83.9192	42.89571	124.9427	15.8733	151.9651	40.6952	127.1432
122.3377	98.51239	146.16298	63.0505	181.6248	97.2344	147.441
148.1364	114.11431	182.15859	84.0675	212.2054	112.2894	183.9835
169.3318	134.4239	204.23972	104.7882	233.8755	132.5514	206.1122
129.2326	93.95954	164.50566	64.4907	193.9745	92.0675	166.3977
166.3231	131.23185	201.41433	101.6801	230.9661	129.3496	203.2966

Prototype in R:

```
#REMEMBER TO DOWNLOAD ISwR PACKAGE
#LOAD ISwR PACKAGE
require(ISwR)

#LOAD cystfibr DATA FROM ISwR
data(cystfibr)
attach(cystfibr)

#SPECIFY FULL & REDUCED MODELS
FM=lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc)
RM=lm(pemax~age+sex+weight+bmp)

#FOR REGRESSION COEFFICIENTS (Estimate):
#FOR OVERALL F TEST OF REGRESSION:
#FOR PARTIAL t/F TESTS OF SINGLE COEFFICIENTS:
#FOR R-SQUARED OR R AND ADJUSTED R-SQUARED
summary(FM)
```

```
> summary(FM)
Call:

lm(formula = pemax ~ age + sex + height + weight + bmp + fev1 +
    rv + frc + tlc)

Residuals:
    Min      1Q  Median      3Q     Max
-37.338 -11.532   1.081  13.386  33.405
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 176.0582   225.8912   0.779    0.448
age          -2.5420     4.8017  -0.529    0.604
sex          -3.7368    15.4598  -0.242    0.812
height       -0.4463     0.9034  -0.494    0.628
weight        2.9928     2.0080   1.490    0.157
bmp          -1.7449     1.1552  -1.510    0.152
fev1          1.0807     1.0809   1.000    0.333
rv            0.1970     0.1962   1.004    0.331
frc          -0.3084     0.4924  -0.626    0.540
tlc           0.1886     0.4997   0.377    0.711

Residual standard error: 25.47 on 15 degrees of freedom
Multiple R-Squared: 0.6373,     Adjusted R-squared: 0.4197
F-statistic: 2.929 on 9 and 15 DF,  p-value: 0.03195
```

#FOR ANOVA TABLE INCLUDING PARTIAL SUMS OF SQUARES
#THAT ADD TO THE REGRESSION SUM OF SQUARES
#FOR RESIDUAL/ERROR SUM OF SQuARES:
#FOR F TESTS OF SERIALLY ADDED INDEPENDENT VARIABLES:
anova(FM)

> anova(FM)

```
Analysis of Variance Table
Response: pemax
          Df  Sum Sq Mean Sq F value   Pr(>F)
age        1 10098.5 10098.5 15.5661 0.001296 **
sex        1   955.4   955.4  1.4727 0.243680
height     1   155.0   155.0  0.2389 0.632089
weight     1   632.3   632.3  0.9747 0.339170
bmp        1  2862.2  2862.2  4.4119 0.053010 .
fev1       1  1549.1  1549.1  2.3878 0.143120
rv         1   561.9   561.9  0.8662 0.366757
frc        1   194.6   194.6  0.2999 0.592007
tlc        1    92.4    92.4  0.1424 0.711160
Residuals 15  9731.2   648.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#FOR GLM TEST OF FULL VS REDUCED MODELS:
anova(FM,RM)

> anova(FM,RM)

```
Analysis of Variance Table
Model 1: pemax ~ age + sex + height + weight + bmp + fev1 + rv +
frc + tlc
Model 2: pemax ~ age + sex + weight + bmp

  Res.Df     RSS Df Sum of Sq     F Pr(>F)
1     15  9731.2
2     20 12602.1 -5   -2870.9 0.885 0.5149
```

#FOR CONFIDENCE AND PREDICTION INTERVALS:

predict(FM,interval="confidence",level=0.95)

predict(FM,interval="prediction",level=0.95)

> predict(FM,interval="confidence",level=0.95)

```
          fit       lwr      upr
1   84.96898  43.71943 126.2185
2   88.41376  52.67346 124.1540
3   86.61406  51.30451 121.9236
4   96.53192  59.76620 133.2976
5   76.30858  35.75649 116.8607
6  111.55172  79.83738 143.2661
7   76.48024  46.34711 106.6134
8   89.96630  56.66590 123.2667
9   90.30724  65.22697 115.3875
10 108.18220  71.73172 144.6327
11  94.35363  65.36953 123.3377
12  79.25203  44.47304 114.0310
13 103.66412  78.88152 128.4467
14 113.11774  76.51811 149.7174
15 123.53963  91.82311 155.2561
16 100.59483  65.89347 135.2962
17 143.96636 111.34213 176.5906
18 123.00215  81.63318 164.3711
19 117.90369  83.55925 152.2481
20  83.91920  42.89571 124.9427
21 122.33769  98.51239 146.1630
22 148.13645 114.11431 182.1586
23 169.33181 134.42390 204.2397
24 129.23260  93.95954 164.5057
25 166.32309 131.23185 201.4143
```

> predict(FM,interval="prediction",level=0.95)

```
          fit        lwr      upr
1   84.96898  16.786531 153.1514
2   88.41376  23.416144 153.4114
3   86.61406  21.852298 151.3758
4   96.53192  30.964860 162.0990
5   76.30858   8.545805 144.0713
6  111.55172  48.677870 174.4256
7   76.48024  14.388963 138.5715
8   89.96630  26.277698 153.6549
9   90.30724  30.504722 150.1098
10 108.18220  42.791386 173.5730
11  94.35363  32.811781 155.8955
12  79.25203  14.778012 143.7261
13 103.66412  43.985827 163.3424
14 113.11774  47.643666 178.5918
15 123.53963  60.664682 186.4146
16 100.59483  36.162649 165.0270
17 143.96636  80.628682 207.3040
18 123.00215  54.747388 191.2569
19 117.90369  53.663036 182.1443
20  83.91920  15.873276 151.9651
21 122.33769  63.050539 181.6248
22 148.13645  84.067522 212.2054
23 169.33181 104.788154 233.8755
24 129.23260  64.490727 193.9745
25 166.32309 101.680099 230.9661
Warning message:
Predictions on current data refer to
_future_ responses in: predict.lm(FM,
interval = "prediction", level = 0.95)
```