

Chapter 6: Variable Screening Methods

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University



Variable Screening Methods

Introduction: Why Use a Variable Screening Method?

Researches often will collect a data set with a **large** number of independent variables, each of which is a potential predictor of some dependent variable, y .

The problem of deciding to include multiple regression model for $E(y)$ is common. Suppose y depends on 10 x 's. 7 quantitative and 3 qualitative yield 288 terms.

$$\begin{aligned}
 E(y) = & \beta_0 + \boxed{\beta_1 x_1 + \dots + \beta_7 x_7} + \beta_8 x_1 x_2 + \dots + \beta_{28} x_6 x_7 \\
 & + \boxed{\beta_{29} x_1^2 + \dots + \beta_{35} x_7^2} + \boxed{\beta_{36} x_8 + \dots + \beta_{38} x_{10}} \\
 & + \beta_{39} x_8 x_9 + \dots + \beta_{42} x_8 x_9 x_{10} + \beta_{43} x_1 x_8 + \dots + \beta_{77} x_7^2 x_8 \\
 & + \beta_{78} x_1 x_9 + \dots + \beta_{112} x_7^2 x_9 + \dots + \beta_{113} x_1 x_{10} + \dots + \beta_{147} x_7^2 x_{10} \\
 & + \beta_{148} x_1 x_8 x_9 + \dots + \beta_{182} x_7^2 x_8 x_9 + \beta_{183} x_1 x_8 x_{10} + \dots + \beta_{217} x_7^2 x_8 x_{10} \\
 & + \beta_{218} x_1 x_9 x_{10} + \dots + \beta_{252} x_7^2 x_9 x_{10} + \beta_{253} x_1 x_8 x_9 x_{10} + \dots + \beta_{287} x_7^2 x_8 x_9 x_{10}
 \end{aligned}$$

Too complex to be
practicably useful.

Variable Screening Methods

Stepwise Regression (Forward Selection)

Stepwise Regression: The user identifies the set of potentially important independent variables x 's that influence the dependent (response) variable y .

Step 1: Fit all possible one-variable models of the form $E(y)=\beta_0+\beta_1x_i, i=1, \dots, k$.

Perform the t -test $H_0: \beta_1=0$ vs. $H_a: \beta_1 \neq 0$.

$t = \hat{\beta}_i / s\sqrt{W_{ii}}$, W_{ii} is the i^{th} diagonal element of $W=(X'X)^{-1}$.

Select the best one variable model (largest $|t|$ statistic). Call it x_1

Step 2: Fit all two variable models with remaining x 's, $E(y)=\beta_0+\beta_1x_1+\beta_2x_i, i \neq 1$.

Perform the t -test $H_0: \beta_2=0$ vs. $H_a: \beta_2 \neq 0$.

$t = \hat{\beta}_i / s\sqrt{W_{ii}}$, W_{ii} is the i^{th} diagonal element of $W=(X'X)^{-1}$.

Select the best two variable model (largest $|t|$ statistic). Call it x_2

Go back and check the t -value of $\hat{\beta}_1$ after $\hat{\beta}_2$ has been added to the model.

Variable Screening Methods

Stepwise Regression (Forward Selection)

Step 3: Fit all three variable models with remaining x 's, $E(y)=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_i$, $i\neq 1,2$.

Perform the t -test $H_0: \beta_3=0$ vs. $H_a: \beta_3\neq 0$.

$t = \hat{\beta}_i / s\sqrt{W_{ii}}$, W_{ii} is the i^{th} diagonal element of $W=(X'X)^{-1}$.

Select the best two variable model (largest $|t|$ statistic). Call it x_2

Go back and check the t -values of $\hat{\beta}_1, \hat{\beta}_2$ after $\hat{\beta}_3$ has been added.

This procedure is continued until no further independent variables can be found that yield significant t -values (at the specified α level) in the presence of the variables already in the model.

Variable Screening Methods Stepwise Regression (Forward Selection)

Example: Log salary y depends on 7 quantitative and 3 qualitative x variables. Which (linear) variables are important.

Independent Variable	Description
x_1	Experience (years)—quantitative
x_2	Education (years)—quantitative
x_3	Gender (1 if male, 0 if female)—qualitative
x_4	Number of employees supervised—quantitative
x_5	Corporate assets (millions of dollars)—quantitative
x_6	Board member (1 if yes, 0 if no)—qualitative
x_7	Age (years)—quantitative
x_8	Company profits (past 12 months, millions of dollars)—quantitative
x_9	Has international responsibility (1 if yes, 0 if no)—qualitative
x_{10}	Company's total sales (past 12 months, millions of dollars)—quantitative

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	11.091	.033		335.524	.000
	X1	.028	.002	.787	12.618	.000
2	(Constant)	10.968	.032		342.659	.000
	X1	.027	.002	.770	15.134	.000
	X3	.197	.028	.361	7.097	.000
3	(Constant)	10.783	.036		298.170	.000
	X1	.027	.001	.771	18.801	.000
	X3	.233	.023	.427	10.170	.000
	X4	.000	.000	.307	7.323	.000
4	(Constant)	10.278	.066		155.154	.000
	X1	.027	.001	.771	24.677	.000
	X3	.232	.017	.425	13.297	.000
	X4	.001	.000	.354	10.920	.000
	X2	.030	.004	.266	8.379	.000
5	(Constant)	9.962	.101		98.578	.000
	X1	.027	.001	.771	26.501	.000
	X3	.225	.016	.412	13.742	.000
	X4	.001	.000	.337	11.064	.000
	X2	.029	.003	.258	8.719	.000
	X5	.002	.000	.116	3.947	.000

R Code
Output
→
t-statistics

Variable Screening Methods

Stepwise Regression (Forward Selection)

<p>lm(formula = y ~ x1, data = df)</p> <p>Coefficients:</p> <table border="1"> <thead> <tr> <th></th> <th>Estimate</th> <th>Std. Error</th> <th>t value</th> <th>Pr(> t)</th> </tr> </thead> <tbody> <tr> <td>(Intercept)</td> <td>11.090887</td> <td>0.033055</td> <td>335.52</td> <td><2e-16 ***</td> </tr> <tr> <td>x1</td> <td>0.027839</td> <td>0.002206</td> <td>12.62</td> <td><2e-16 ***</td> </tr> </tbody> </table> <p>Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 0.1612 on 98 degrees of freedom</p> <p>Multiple R-squared: 0.619, Adjusted R-squared: 0.6151</p> <p>F-statistic: 159.2 on 1 and 98 DF, p-value: <2.2e-16</p>		Estimate	Std. Error	t value	Pr(> t)	(Intercept)	11.090887	0.033055	335.52	<2e-16 ***	x1	0.027839	0.002206	12.62	<2e-16 ***	<p>lm(formula = y ~ x2, data = df)</p> <p>Coefficients:</p> <table border="1"> <thead> <tr> <th></th> <th>Estimate</th> <th>Std. Error</th> <th>t value</th> <th>Pr(> t)</th> </tr> </thead> <tbody> <tr> <td>(Intercept)</td> <td>11.05594</td> <td>0.17971</td> <td>61.520</td> <td><2e-16 ***</td> </tr> <tr> <td>x2</td> <td>0.02491</td> <td>0.01110</td> <td>2.243</td> <td>0.0271 *</td> </tr> </tbody> </table> <p>Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 0.2547 on 98 degrees of freedom</p> <p>Multiple R-squared: 0.04884, Adjusted R-squared: 0.03914</p> <p>F-statistic: 5.032 on 1 and 98 DF, p-value: 0.02713</p>		Estimate	Std. Error	t value	Pr(> t)	(Intercept)	11.05594	0.17971	61.520	<2e-16 ***	x2	0.02491	0.01110	2.243	0.0271 *	<p>lm(formula = y ~ x3, data = df)</p> <p>Coefficients:</p> <table border="1"> <thead> <tr> <th></th> <th>Estimate</th> <th>Std. Error</th> <th>t value</th> <th>Pr(> t)</th> </tr> </thead> <tbody> <tr> <td>(Intercept)</td> <td>11.31231</td> <td>0.04112</td> <td>275.116</td> <td><2e-16 ***</td> </tr> <tr> <td>x3</td> <td>0.21623</td> <td>0.05061</td> <td>4.272</td> <td>4.49e-05 ***</td> </tr> </tbody> </table> <p>Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 0.2398 on 98 degrees of freedom</p> <p>Multiple R-squared: 0.157, Adjusted R-squared: 0.1484</p> <p>F-statistic: 18.25 on 1 and 98 DF, p-value: 4.487e-05</p>		Estimate	Std. Error	t value	Pr(> t)	(Intercept)	11.31231	0.04112	275.116	<2e-16 ***	x3	0.21623	0.05061	4.272	4.49e-05 ***	<p>lm(formula = y ~ x4, data = df)</p> <p>Coefficients:</p> <table border="1"> <thead> <tr> <th></th> <th>Estimate</th> <th>Std. Error</th> <th>t value</th> <th>Pr(> t)</th> </tr> </thead> <tbody> <tr> <td>(Intercept)</td> <td>1.134e+01</td> <td>5.813e-02</td> <td>195.157</td> <td><2e-16 ***</td> </tr> <tr> <td>x4</td> <td>3.236e-04</td> <td>1.535e-04</td> <td>2.107</td> <td>0.0376 *</td> </tr> </tbody> </table> <p>Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 0.2554 on 98 degrees of freedom</p> <p>Multiple R-squared: 0.04335, Adjusted R-squared: 0.03359</p> <p>F-statistic: 4.441 on 1 and 98 DF, p-value: 0.03763</p>		Estimate	Std. Error	t value	Pr(> t)	(Intercept)	1.134e+01	5.813e-02	195.157	<2e-16 ***	x4	3.236e-04	1.535e-04	2.107	0.0376 *	<p>Call: lm(formula = y ~ x5, data = df)</p> <p>Coefficients:</p> <table border="1"> <thead> <tr> <th></th> <th>Estimate</th> <th>Std. Error</th> <th>t value</th> <th>Pr(> t)</th> </tr> </thead> <tbody> <tr> <td>(Intercept)</td> <td>10.853365</td> <td>0.293139</td> <td>37.02</td> <td><2e-16 ***</td> </tr> <tr> <td>x5</td> <td>0.003436</td> <td>0.001668</td> <td>2.06</td> <td>0.042 *</td> </tr> </tbody> </table> <p>Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 0.2557 on 98 degrees of freedom</p> <p>Multiple R-squared: 0.04152, Adjusted R-squared: 0.03174</p> <p>F-statistic: 4.245 on 1 and 98 DF, p-value: 0.04202</p>		Estimate	Std. Error	t value	Pr(> t)	(Intercept)	10.853365	0.293139	37.02	<2e-16 ***	x5	0.003436	0.001668	2.06	0.042 *
	Estimate	Std. Error	t value	Pr(> t)																																																																											
(Intercept)	11.090887	0.033055	335.52	<2e-16 ***																																																																											
x1	0.027839	0.002206	12.62	<2e-16 ***																																																																											
	Estimate	Std. Error	t value	Pr(> t)																																																																											
(Intercept)	11.05594	0.17971	61.520	<2e-16 ***																																																																											
x2	0.02491	0.01110	2.243	0.0271 *																																																																											
	Estimate	Std. Error	t value	Pr(> t)																																																																											
(Intercept)	11.31231	0.04112	275.116	<2e-16 ***																																																																											
x3	0.21623	0.05061	4.272	4.49e-05 ***																																																																											
	Estimate	Std. Error	t value	Pr(> t)																																																																											
(Intercept)	1.134e+01	5.813e-02	195.157	<2e-16 ***																																																																											
x4	3.236e-04	1.535e-04	2.107	0.0376 *																																																																											
	Estimate	Std. Error	t value	Pr(> t)																																																																											
(Intercept)	10.853365	0.293139	37.02	<2e-16 ***																																																																											
x5	0.003436	0.001668	2.06	0.042 *																																																																											
<p>lm(formula = y ~ x6, data = df)</p> <p>Coefficients:</p> <table border="1"> <thead> <tr> <th></th> <th>Estimate</th> <th>Std. Error</th> <th>t value</th> <th>Pr(> t)</th> </tr> </thead> <tbody> <tr> <td>(Intercept)</td> <td>11.46777</td> <td>0.03652</td> <td>314.017</td> <td><2e-16 ***</td> </tr> <tr> <td>x6</td> <td>-0.02603</td> <td>0.05217</td> <td>-0.499</td> <td>0.619</td> </tr> </tbody> </table> <p>Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 0.2608 on 98 degrees of freedom</p> <p>Multiple R-squared: 0.002533, Adjusted R-squared: -0.007645</p> <p>F-statistic: 0.2489 on 1 and 98 DF, p-value: 0.619</p>		Estimate	Std. Error	t value	Pr(> t)	(Intercept)	11.46777	0.03652	314.017	<2e-16 ***	x6	-0.02603	0.05217	-0.499	0.619	<p>lm(formula = y ~ x7, data = df)</p> <p>Coefficients:</p> <table border="1"> <thead> <tr> <th></th> <th>Estimate</th> <th>Std. Error</th> <th>t value</th> <th>Pr(> t)</th> </tr> </thead> <tbody> <tr> <td>(Intercept)</td> <td>10.682669</td> <td>0.098393</td> <td>108.571</td> <td><2e-16 ***</td> </tr> <tr> <td>x7</td> <td>0.018029</td> <td>0.002247</td> <td>8.022</td> <td>2.28e-12 ***</td> </tr> </tbody> </table> <p>Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 0.2029 on 98 degrees of freedom</p> <p>Multiple R-squared: 0.3964, Adjusted R-squared: 0.3902</p> <p>F-statistic: 64.35 on 1 and 98 DF, p-value: 2.277e-12</p>		Estimate	Std. Error	t value	Pr(> t)	(Intercept)	10.682669	0.098393	108.571	<2e-16 ***	x7	0.018029	0.002247	8.022	2.28e-12 ***	<p>lm(formula = y ~ x8, data = df)</p> <p>Coefficients:</p> <table border="1"> <thead> <tr> <th></th> <th>Estimate</th> <th>Std. Error</th> <th>t value</th> <th>Pr(> t)</th> </tr> </thead> <tbody> <tr> <td>(Intercept)</td> <td>11.388078</td> <td>0.132479</td> <td>85.961</td> <td><2e-16 ***</td> </tr> <tr> <td>x8</td> <td>0.008693</td> <td>0.016868</td> <td>0.515</td> <td>0.607</td> </tr> </tbody> </table> <p>Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 0.2608 on 98 degrees of freedom</p> <p>Multiple R-squared: 0.002703, Adjusted R-squared: -0.007474</p> <p>F-statistic: 0.2656 on 1 and 98 DF, p-value: 0.6075</p>		Estimate	Std. Error	t value	Pr(> t)	(Intercept)	11.388078	0.132479	85.961	<2e-16 ***	x8	0.008693	0.016868	0.515	0.607	<p>lm(formula = y ~ x9, data = df)</p> <p>Coefficients:</p> <table border="1"> <thead> <tr> <th></th> <th>Estimate</th> <th>Std. Error</th> <th>t value</th> <th>Pr(> t)</th> </tr> </thead> <tbody> <tr> <td>(Intercept)</td> <td>11.45432</td> <td>0.02884</td> <td>397.211</td> <td><2e-16 ***</td> </tr> <tr> <td>x9</td> <td>0.00386</td> <td>0.06797</td> <td>0.057</td> <td>0.955</td> </tr> </tbody> </table> <p>Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 0.2611 on 98 degrees of freedom</p> <p>Multiple R-squared: 3.291e-05, Adjusted R-squared: -0.01017</p> <p>F-statistic: 0.003225 on 1 and 98 DF, p-value: 0.9548</p>		Estimate	Std. Error	t value	Pr(> t)	(Intercept)	11.45432	0.02884	397.211	<2e-16 ***	x9	0.00386	0.06797	0.057	0.955	<p>lm(formula = y ~ x10, data = df)</p> <p>Coefficients:</p> <table border="1"> <thead> <tr> <th></th> <th>Estimate</th> <th>Std. Error</th> <th>t value</th> <th>Pr(> t)</th> </tr> </thead> <tbody> <tr> <td>(Intercept)</td> <td>11.325878</td> <td>0.238765</td> <td>47.435</td> <td><2e-16 ***</td> </tr> <tr> <td>x10</td> <td>0.005201</td> <td>0.009558</td> <td>0.544</td> <td>0.588</td> </tr> </tbody> </table> <p>Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 0.2607 on 98 degrees of freedom</p> <p>Multiple R-squared: 0.003012, Adjusted R-squared: -0.007161</p> <p>F-statistic: 0.2961 on 1 and 98 DF, p-value: 0.5876</p>		Estimate	Std. Error	t value	Pr(> t)	(Intercept)	11.325878	0.238765	47.435	<2e-16 ***	x10	0.005201	0.009558	0.544	0.588
	Estimate	Std. Error	t value	Pr(> t)																																																																											
(Intercept)	11.46777	0.03652	314.017	<2e-16 ***																																																																											
x6	-0.02603	0.05217	-0.499	0.619																																																																											
	Estimate	Std. Error	t value	Pr(> t)																																																																											
(Intercept)	10.682669	0.098393	108.571	<2e-16 ***																																																																											
x7	0.018029	0.002247	8.022	2.28e-12 ***																																																																											
	Estimate	Std. Error	t value	Pr(> t)																																																																											
(Intercept)	11.388078	0.132479	85.961	<2e-16 ***																																																																											
x8	0.008693	0.016868	0.515	0.607																																																																											
	Estimate	Std. Error	t value	Pr(> t)																																																																											
(Intercept)	11.45432	0.02884	397.211	<2e-16 ***																																																																											
x9	0.00386	0.06797	0.057	0.955																																																																											
	Estimate	Std. Error	t value	Pr(> t)																																																																											
(Intercept)	11.325878	0.238765	47.435	<2e-16 ***																																																																											
x10	0.005201	0.009558	0.544	0.588																																																																											

Variable Screening Methods

Stepwise Regression Most look at R^2 or R_a^2 instead of t -statistic.

Index	N	Predictors	R-Square	Adj. R-Square	Mallow's Cp	Index	N	Predictors	R-Square	Adj. R-Square	Mallow's Cp
1	1	x1	0.6189794572	0.615091492	343.85658	46	26	x6 x7	0.3974306886	0.385006579	601.61624
7	2	x7	0.3963753117	0.390215876	600.83459	51	27	x7 x9	0.3973609447	0.384935397	601.69676
3	3	x3	0.1570025883	0.148400574	877.17052	28	28	x3 x4	0.2466304387	0.231097046	775.70261
2	4	x2	0.0488408794	0.039135174	1002.03423	20	29	x2 x3	0.1986675946	0.182145277	831.07173
4	5	x4	0.0433532791	0.033591578	1008.36921	29	30	x3 x5	0.1858378993	0.169051052	845.88255
5	6	x5	0.0415163660	0.031735921	1010.48978	34	31	x3 x10	0.1722383501	0.155171100	861.58210
10	7	x10	0.0030120259	-0.007161321	1054.93984	30	32	x3 x6	0.1661114504	0.148917872	868.65510
8	8	x8	0.0027029290	-0.007473572	1055.29667	32	33	x3 x8	0.1583617684	0.141008403	877.60146
6	9	x6	0.0025329167	-0.007645319	1055.49293	33	34	x3 x9	0.1576238634	0.140255283	878.45331
9	10	x9	0.0000329115	-0.010170834	1058.37898	21	35	x2 x4	0.1123065965	0.094003640	930.76833
12	11	x1 x3	0.7492074614	0.744036481	195.51916	22	36	x2 x5	0.0864422276	0.067605985	960.62660
11	12	x1 x2	0.6676461635	0.660793507	289.67491	35	37	x4 x5	0.0766236732	0.057584986	971.96131
13	13	x1 x4	0.6657473331	0.658855526	291.86695	27	38	x2 x10	0.0510848803	0.031519620	1001.44372
14	14	x1 x5	0.6591000500	0.652071185	299.54069	25	39	x2 x8	0.0503306478	0.030749836	1002.31442
15	15	x1 x6	0.6258493043	0.618134857	337.92591	23	40	x2 x6	0.0502055536	0.030622163	1002.45883
19	16	x1 x10	0.6251626373	0.617434032	338.71861	26	41	x2 x9	0.0489118938	0.029301830	1003.95225
17	17	x1 x8	0.6211730639	0.613362199	343.32425	40	42	x4 x10	0.0472231622	0.027578279	1005.90175
16	18	x1 x7	0.6202405195	0.612410427	344.40079	38	43	x4 x8	0.0460452751	0.026376105	1007.26153
18	19	x1 x9	0.6195497268	0.611705391	345.19825	45	44	x5 x10	0.0437062688	0.023988872	1009.96172
31	20	x3 x7	0.5097037383	0.499594537	472.00633	39	45	x4 x9	0.0433907498	0.023666848	1010.32596
24	21	x2 x7	0.4647957052	0.453760565	523.84892	36	46	x4 x6	0.0433809373	0.023656833	1010.33728
42	22	x5 x7	0.4466738810	0.435265095	544.76906	44	47	x5 x9	0.0421443980	0.022394798	1011.76477
37	23	x4 x7	0.4294675355	0.417703979	564.63236	41	48	x5 x6	0.0419712177	0.022218047	1011.96469
52	24	x7 x10	0.3992410018	0.386854218	599.52639	43	49	x5 x8	0.0417417885	0.021983887	1012.22955
50	25	x7 x8	0.3976158033	0.385195511	601.40254	54	50	x8 x10	0.0057435143	-0.014756619	1053.78656
						49	51	x6 x10	0.0052147962	-0.015296239	1054.39693

Variable Screening Methods

Stepwise Regression

```

# R Code
# read data
mydata<-read.delim("execsal.txt",header=FALSE,sep=" ",dec=".")
# parse out variables
n <- nrow(mydata)
k <- ncol(mydata)-1

# Parse all variables
y <- c(mydata[, 1]) #ln salary
x1 <- c(mydata[, 2]) #x1
x2 <- c(mydata[, 3]) #x2
x3 <- c(mydata[, 4]) #x3
x4 <- c(mydata[, 5]) #x4
x5 <- c(mydata[, 6]) #x5
x6 <- c(mydata[, 7]) #x6
x7 <- c(mydata[, 8]) #x7
x8 <- c(mydata[, 9]) #x8
x9 <- c(mydata[,10]) #x9
x10<- c(mydata[,11]) #x10
df<- data.frame(cbind(x1,x2,x3,x4,x5,x6,x7,x8,x9,x10))
names(df)<-c("x1","x2","x3","x4","x5","x6","x7","x8","x9","x10")

# one at a time fits
lmx1 <- lm(y~x1,data=df)
summary.lm(lmx1)
lmx2 <- lm(y~x2,data=df)
summary.lm(lmx2)
lmx3 <- lm(y~x3,data=df)
summary.lm(lmx3)
lmx4 <- lm(y~x4,data=df)
summary.lm(lmx4)
lmx5 <- lm(y~x5,data=df)
summary.lm(lmx5)
lmx6 <- lm(y~x6,data=df)
summary.lm(lmx6)
lmx7 <- lm(y~x7,data=df)
summary.lm(lmx7)
lmx8 <- lm(y~x8,data=df)
summary.lm(lmx8)
lmx9 <- lm(y~x9,data=df)
summary.lm(lmx9)
lmx10 <- lm(y~x10,data=df)
summary.lm(lmx10)

# use stepwise function
install.packages("olsrr")
library(olsrr)
model = lm(y~.,data=df)
k=ols_step_all_possible(model,max_order=3)
k

```

Variable Screening Methods

All-Possible-Regressions Selection Procedure

There are several criteria that can be used.

$$1. R^2 = 1 - \frac{SSE}{SS(Total)}$$

R-Square

$$2. R_a^2 = 1 - (n - 1) \left[\frac{MSE}{SS(Total)} \right]$$

Adjusted R-Square

$$3. C_p = \frac{SSE_p}{MSE_k} + 2(p + 1) - n$$

Mallow's C_p

$$4. PRESS = \sum_{i=1}^n [y_i - \hat{y}_{(i)}]^2$$

Predictive Sum of Squares

1. R^2 criterion

Looking for a simple model that is as good as, or nearly as good as, the model with all k independent variables.

2. Adjusted R^2 or MSE criterion

Prefer the model with largest, or near largest, adjusted R^2 .

3. Mallow's C_p Criterion

Prefer a small value of C_p and a value of C_p near $p+1$.

4. *PRESS* Criterion

Desire a model with a small *PRESS*.

Variable Screening Methods

All-Possible-Regressions Selection Procedure

There are several criteria that can be used.

$$1. R^2 = 1 - \frac{SSE}{SS(Total)}$$

R-Square

$$2. R_a^2 = 1 - (n - 1) \left[\frac{MSE}{SS(Total)} \right]$$

Adjusted R-Square

$$3. C_p = \frac{SSE_p}{MSE_k} + 2(p + 1) - n$$

Mallow's C_p

$$4. PRESS = \sum_{i=1}^n [y_i - \hat{y}_{(i)}]^2$$

Predictive Sum of Squares

Number of Predictors p	Variables in the Model	R^2	adj- R^2	MSE	C_p	PRESS
1	x_1	.619	.615	.0260	343.9	2.664
2	x_1, x_3	.749	.744	.0173	195.5	1.788
3	x_1, x_3, x_4	.839	.834	.0112	93.8	1.171
4	x_1, x_2, x_3, x_4	.907	.904	.0065	16.8	.696
5	x_1, x_2, x_3, x_4, x_5	.921	.916	.0056	3.6	.610
6	$x_1, x_2, x_3, x_4, x_5, x_9$.922	.917	.0056	4.0	.610
7	$x_1, x_2, x_3, x_4, x_5, x_6, x_9$.923	.917	.0056	5.4	.620
8	$x_1, x_2, x_3, x_4, x_5, x_6, x_8, x_9$.923	.916	.0057	7.2	.629
9	$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9$.923	.915	.0057	9.1	.643
10	$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$.923	.914	.0058	11.0	.654

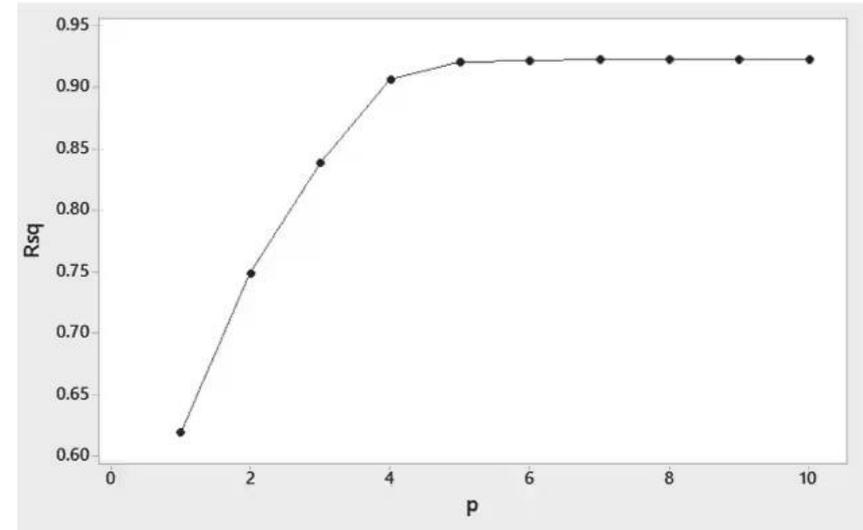
The best one-variable model includes x_1 (years of experience), the best two-variable model includes x_1 and x_3 (gender), the best three-variable model includes x_1, x_3, x_4 (number supervised) and so on.

Variable Screening Methods

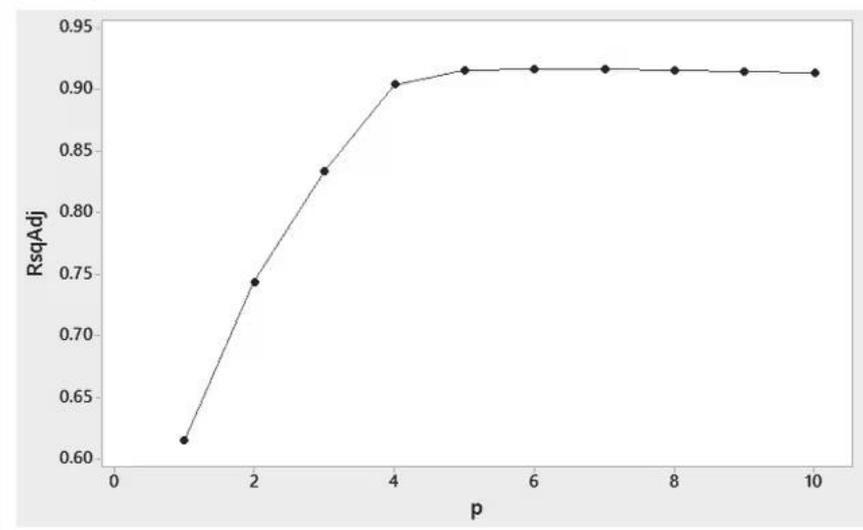
All-Possible-Regressions Selection Procedure

Instead of t statistic, most look at R^2 or R_a^2 .

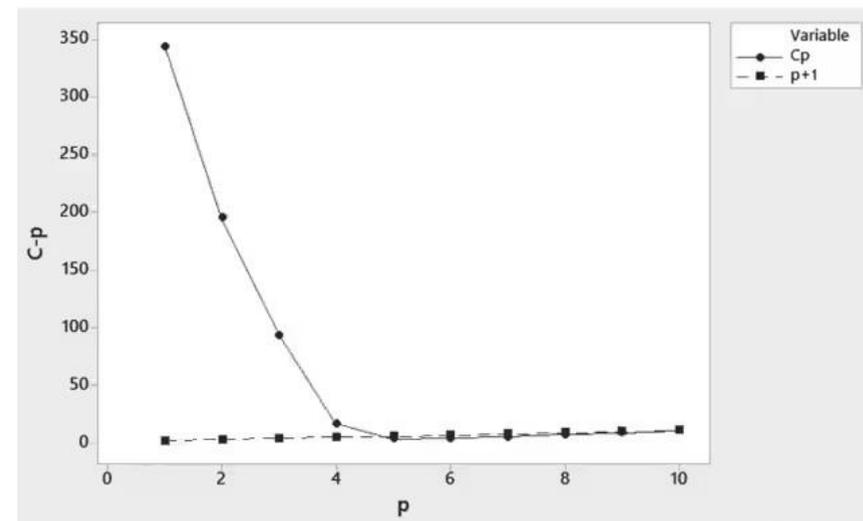
a. R^2 criterion



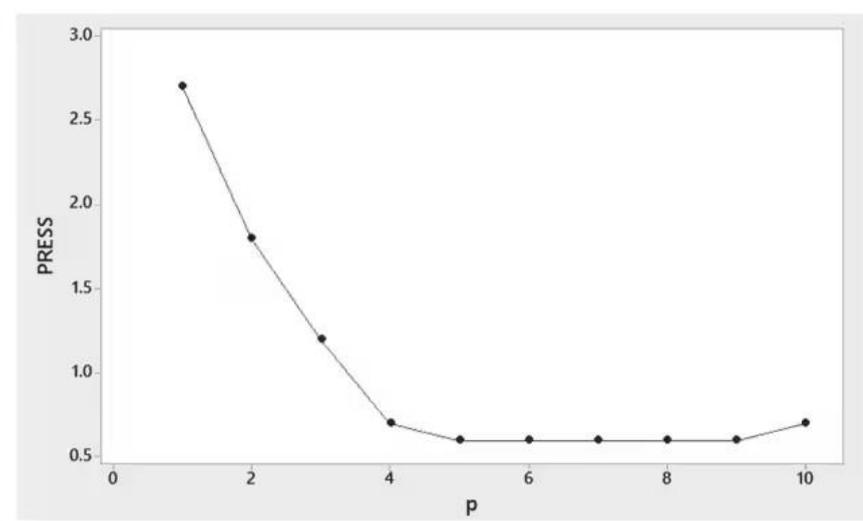
b. R_{adj}^2 criterion



c. C_p criterion



d. PRESS criterion



According to all four criteria, the variables $x_1, x_2, x_3, x_4,$ and x_5 should all be included.

Variable Screening Methods

Homework:

Read Chapter 6

Problems #: 8 (CLERICAL)

Submit at minimum one file with all your answers and another with your code.

Variable Screening Methods

Questions?