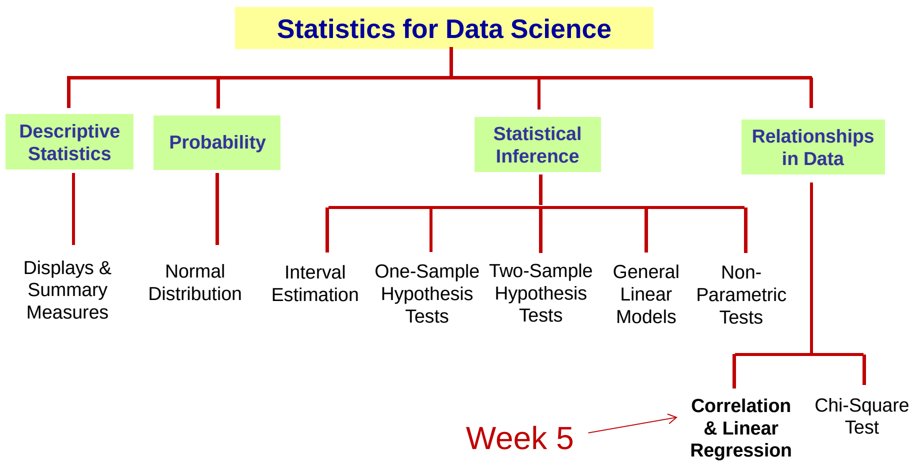
MATH 4044
Statistics for Data Science

Multiple Regression

Course outline



Field, A & Miles, J, *Discovering Statistics Using SAS*, Chapter 7 (sections 7.5-7.11)

Multiple linear regression

- We extend the 'model' part to include more than one explanatory variables

outcome = (model) + error

- For two explanatory variables we have:

$$\hat{y} = \underbrace{b_0 + b_1 x_1 + b_2 x_2}_{\text{model}} + e_i$$

- In general for *p* explanatory variables, we have:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p + e_i$$

Goals of multiple regression

- Describe:
  - Develop a model to describe the relationship between the explanatory variables and the response variable.
- Predict:
  - Use sample data to make predictions of response values from explanatory variables.
- Confirm theories:
  - Which variables, or combination of variables, need to be included in the model?
  - How much does each explanatory variable contribute towards capturing the variability in the response variable?
- Techniques used depend on the objectives of the analysis.

Example: Fitness study

- We have data for 50 subjects based on the following variables:
  - Age of subject (years)
  - Maximum number of push-ups
  - Resting pulse rate (bpm)
  - Maximum pulse rate (bpm)
  - Pulse rate while running (bpm)

- We want to build a model that can predict how many push-ups a person can do.

Example: Correlations

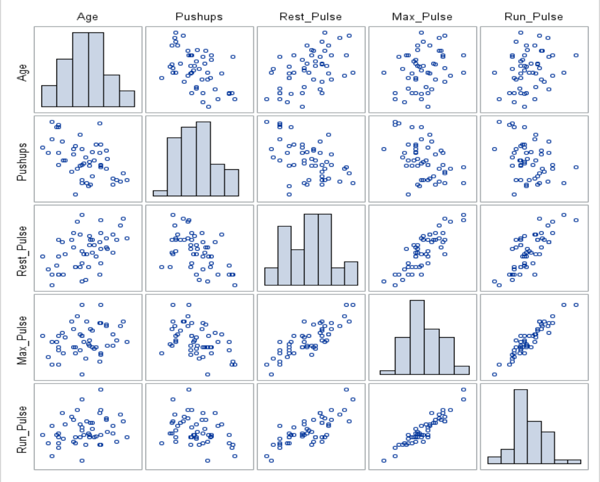
Pearson Correlation Coefficients, N = 50 Prob > |r| under H0: Rho=0	
	Pushups
Age	-0.49191 0.0003
Rest_Pulse	-0.49639 0.0002
Max_Pulse	-0.45010 0.0010
Run_Pulse	-0.34555 0.0140

All four variables are significantly correlated with *Pushups* at 5% significance level.

All four variables are good candidates for explanatory variables.

Example: scatterplot matrix

What are the associations among explanatory variables?



Age is weakly negatively associated with Pushups, and weakly positively associated with all Pulse variables.

Running pulse is weakly positively related to Age, weakly negatively associated with Pushups, but strongly positively related to Rest and Max Pulse.

Example: Overall model fit

- Model 1 that relates *Pushups* to *Age* and *Max_Pulse*:

The REG Procedure
Model: MODEL1
Dependent Variable: Pushups

Number of Observations Read	50
Number of Observations Used	50

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2795.54147	1397.77074	12.75	<.0001
Error	47	5152.95853	109.63742		
Corrected Total	49	7948.50000			

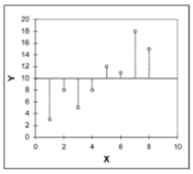
$$F = \frac{MS_M}{MS_R}$$

The model is statistically significant.

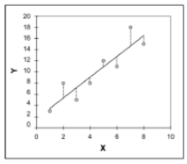
We should reject the null hypothesis that all the betas (except intercept) are zero.

Recall: Sums of squares $SS_T = SS_M + SS_R$

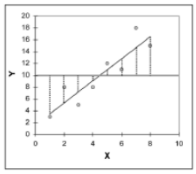
- SS_T
  - Total variability (variability between actual data and the mean)



- SS_R
  - Residual/error variability (variability between the regression model and the actual data)



- SS_M
  - Model variability (difference in variability between the model and the mean)



Significance of a regression model

- Overall test of model adequacy (Analysis of Variance table):

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

$$H_1 : \text{at least one of the coefficients in not zero}$$

- The test statistic is

$$F = \frac{MS_M}{MS_R} = \frac{SS_M / p}{SS_R / (n - p - 1)}$$

p is the number of predictors (excluding the intercept) and *n* is the number of observations

- Note that the intercept, β_0, is not included in the hypotheses.

Example: Coefficient of determination

- Model 1 that relates *Pushups* to *Age* and *Max_Pulse*:

Root MSE	10.47079	R-Square	0.3517
Dependent Mean	25.30000	Adj R-Sq	0.3241
Coeff Var	41.38652		

R^2 = 35.17%

- *Age* and *Max_Pulse* together explain 35% of variability in *Pushups*.
- Use adjusted R^2 to answer the following question:
  - How well does the model generalise?

R^2 vs adjusted R^2

- In multiple regression, R^2 is the square of the multiple correlation coefficient between the dependent variable and the predictors.
- The adjusted R^2 indicates the loss of predictive power or shrinkage.
  - How much variance in y would be accounted for if the model was derived from the population?
  - What is the loss in predictive power?
  - We want this value to be close to our R^2 value.

$$\text{Adjusted } R^2 = 1 - \left(\frac{n-1}{n-p-1}\right)\left(1 - R^2\right)$$

Stein's formula

- How well does the model predict data from a different sample?
  - How well does the model cross-validate?

$$\text{Adjusted } R^2 = 1 - \left[\left(\frac{n-1}{n-p-1}\right)\left(\frac{n-2}{n-p-2}\right)\left(\frac{n+1}{n}\right)\right]\left(1-R^2\right)$$

For our Model 1, this formula gives 0.2806, which is much lower than $R^2 = 0.3517$ so there is room for improvement.

Example: Coefficients

- Model 1 that relates *Pushups* to *Age* and *Max_Pulse*:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	96.65502	18.89988	5.11	<.0001
Age	1	-0.31605	0.09613	-3.29	0.0019
Max_Pulse	1	-0.47750	0.16929	-2.82	0.0070

$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$

Pushups = 96.655 – 0.316 Age – 0.478 Max_Pulse

On average, for a *fixed maximum pulse*, the number of push-ups decreases by 0.316 for each 1 year increase in age.

On average, for a *fixed age*, the number of push-ups decreases by 0.478 for each 1 bpm increase in maximum pulse.

Example: Regression Inference for β_1 and β_2

- Model 1 that relates *Pushups* to *Age* and *Max_Pulse*:

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	96.65502	18.89988	5.11	<.0001
Age	1	-0.31605	0.09613	-3.29	0.0019
Max_Pulse	1	-0.47750	0.16929	-2.82	0.0070

H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$

$t_{48} = -3.29$, p-value = $0.0019 < 0.05$ thus there is a relationship between *Pushups* and *Age*.

H_0: $\beta_2 = 0$
H_1: $\beta_2 \neq 0$

$t_{48} = -2.82$, p-value = $0.007 < 0.05$ thus there is a relationship between *Pushups* and *Max_Pulse*.

Example: Regression Inference for β_1 and β_2

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	96.65502	18.89988	5.11	<.0001	58.63337	134.67666
Age	1	-0.31605	0.09613	-3.29	0.0019	-0.50943	-0.12267
Max_Pulse	1	-0.47750	0.16929	-2.82	0.0070	-0.81807	-0.13693

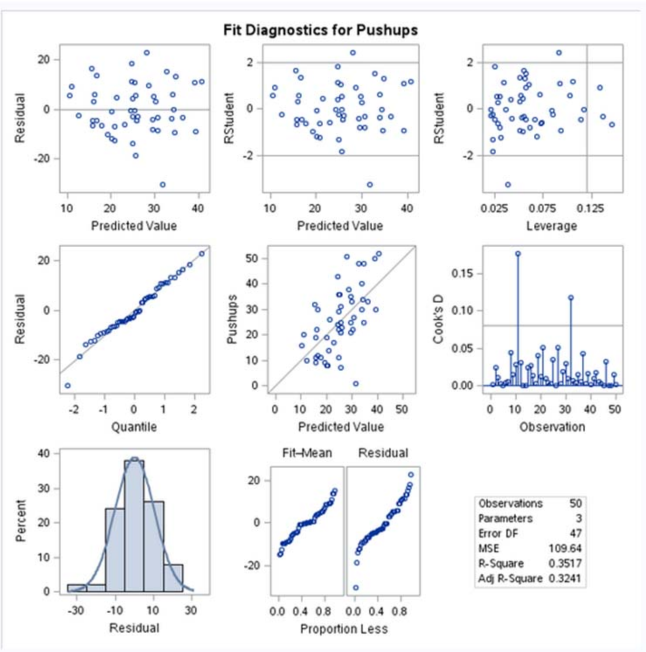
We are 95% confident that the population value of the slope for Age is between -0.509 and -0.123.

We are 95% confident that the population value of the slope for Max_Pulse is between -0.818 and -0.137.

Performing hypothesis tests

- There are no model assumptions needed about the error terms to calculate estimates of the coefficients.
- However, all the model assumptions should be checked before conducting a hypothesis test.

- Assumptions for linear regression:
  - Error terms must be Normally distributed.



Linearity, Independence, Error Variance:

The residual-versus-fitted values plot shows no apparent pattern so linearity and independence are OK.

There is no evidence unequal vertical spread, so constant error variance is also OK.

Normality:

Normal probability plot of residuals shows no curved pattern.

Diagnostic measures - Outliers

- Studentized residuals:
  - Residuals divided by an estimate of their standard deviation.
  - To facilitate interpretation across different models.
  - Cause for concern:
    - Studentized residuals with absolute value greater than 3.
    - More than 1% of sample cases with an absolute value of 2.5.
      The model is a fairly poor fit to the sample data.
    - More than 5% of cases with an absolute value greater than 2.
      The model is a poor representation of the actual data.

Diagnostic measures – Influence

- Adjusted predicted value for a case:
  - Predicted value for a case from a model estimated without that case.
- DFFit:
  - Difference between the adjusted predicted value and the original predicted value.
    - Reported in standardized form.
    - For a non-influential case the value should be zero.
    - Cause for concern: absolute values greater than 1.
    - Rule of thumb: absolute value greater than $2\times\sqrt{\frac{p+1}{n}}$

Diagnostic measures – Influence

- Cook's distance:
  - Measure of overall influence of a case on the model.
    - Impact a case has on the model's ability to predict all cases.
  - Cause for concern: values greater than 1.
  - Rule of thumb: values greater than 4/n.
- PRESS residuals:
  - Differences between adjusted predicted values and original observed values.
    - Influence of case on the ability of the model to predict that case.

Diagnostic measures – Influence

- Studentized deleted residuals (Rstudent in SAS):
  - Standardized values of the PRESS (prediction sum of squares) residuals.
  - PRESS residuals divided by the standard error.
- DFBeta (standardized):
  - Difference between a parameter estimated using all observations and when one observation is excluded.
  - Cause for concern: absolute value greater that 1.
  - Rule of thumb: absolute value greater than $\frac{2}{\sqrt{n}}$

Diagnostic measures – Influence

- Leverage (hat value):
  - Influence of the observed value of the outcome variable over the predicted values.
  - The average leverage is (p+1)/n and values lie between 0 (no influence) and 1 (complete influence).
  - Cause for concern: values greater than 2 or 3 times the average hat value, so greater than

$$\frac{2(p+1)}{n}$$

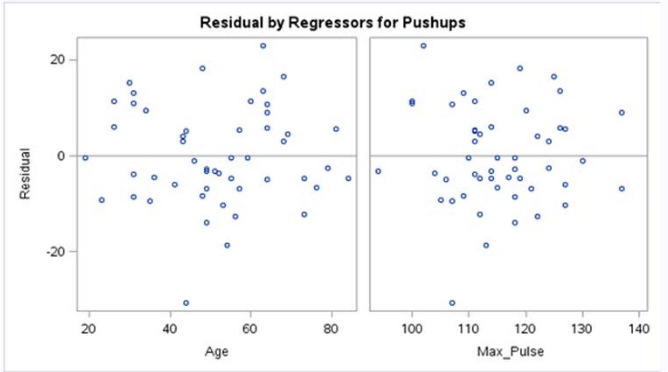
Example: Model diagnostics

Output Statistics

Obs	Subj	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS Intercept	DFBETAS Age	DFBETAS Max_Pulse
1	1	3.0465	0.2958	0.0513	1.1179	0.0687	-0.0365	0.0352	0.0297
2	2	10.6648	1.0520	0.0605	1.0572	0.2670	0.1566	0.1653	-0.1816
3	3	-6.7225	-0.6625	0.0721	1.1172	-0.1846	-0.0145	-0.1569	0.0462
4	4	5.2537	0.5047	0.0271	1.0784	0.0843	0.0406	-0.0248	-0.0275
5	5	-0.3597	-0.0344	0.0213	1.0898	-0.0051	-0.0006	-0.0012	0.0006
6	6	-6.8626	-0.6609	0.0283	1.0670	-0.1128	0.0502	-0.0187	-0.0510
7	7	-4.8127	-0.4717	0.0662	1.1258	-0.1256	-0.0780	-0.0764	0.0896
8	8	15.2615	1.5214	0.0566	0.9758	0.3725	0.0483	-0.2975	0.0452
9	9	-9.5007	-0.9297	0.0503	1.0621	-0.2139	-0.1342	0.1087	0.0921
10	10	-6.7510	-0.6938	0.1460	1.2106	-0.2869	0.2444	0.0858	-0.2660

Diagnostic statistics for the first ten cases

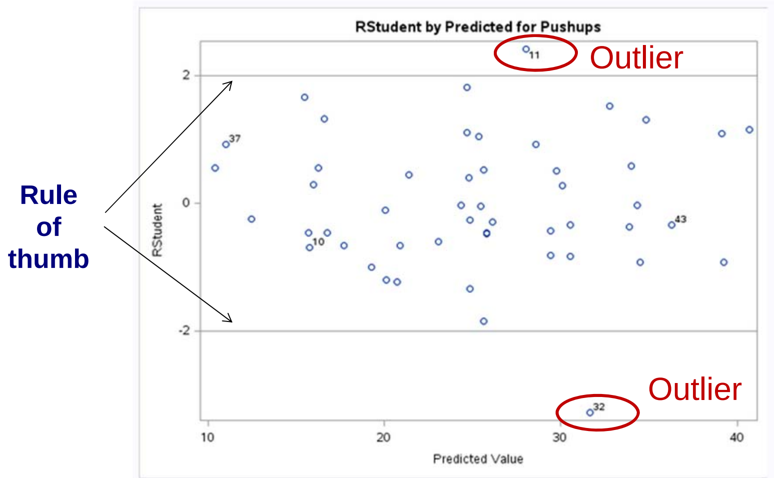
Example: Influence diagnostics



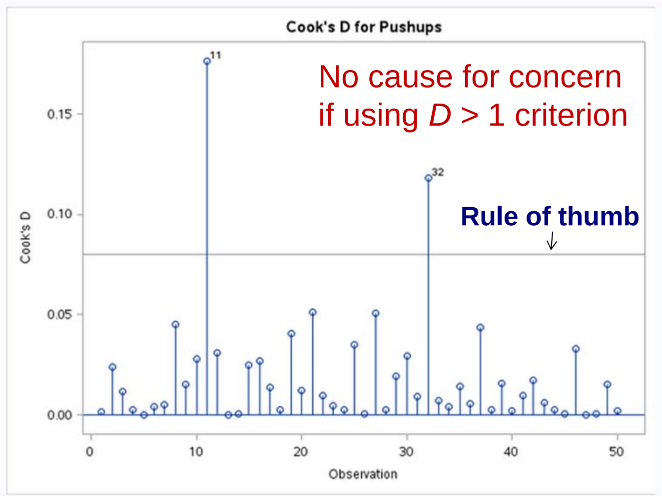
Residuals plotted against each of the predictor variables:

- There does not appear to be any systematic pattern.
- Variability of residuals across values of predictor variables does not seem to show a pattern either.

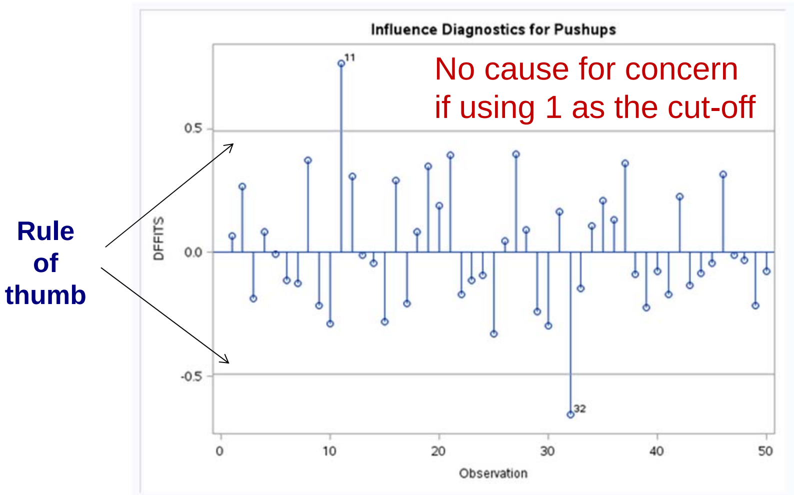
Example: Influence diagnostics



Example: Influence diagnostics

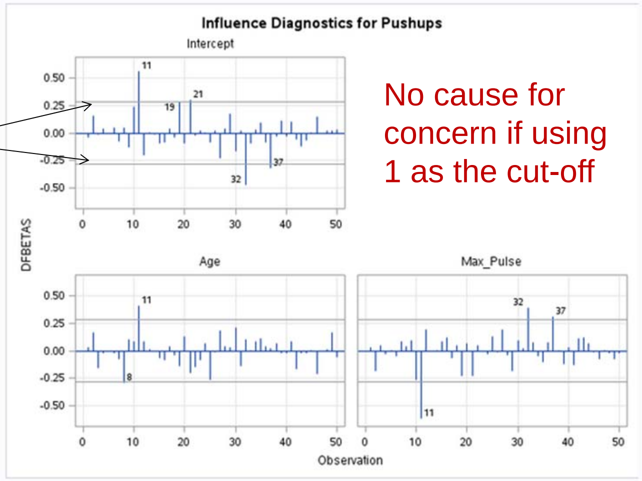


Example: Influence diagnostics



Example: Influence diagnostics

Rule of thumb



No cause for concern if using 1 as the cut-off

Rules of thumb for identifying observations worthy of further investigation

Measure	Value
Studentised residual (absolute value)	> 2
DFFTS (absolute value)	> 2 x sqrt((p+1)/n)
DFBETA (absolute value)	> 2 / sqrt(n)
Leverage	> 2 x (p+1)/n
Cook’s D	> 4 / n

Model selection methods

- R-squared method:
  - Choose the model with highest R^2 out of all possible regression models.
- Mallow's C*p*:
  - Choose the first model in which C*p* is less than or equal to p+1, if the goal is prediction.
  - Hocking: Choose the first model with Cp less than or equal to $2(p+1) - (p_{full} + 1) + 1$, if the goal is to explain relationships.
  - To avoid overfitting.

Model selection methods

- Stepwise methods:
  - Decisions about the order in which predictors enter into the model are based on some mathematical criterion.
- Forward method:
  - Starts with a model based on the intercept only.
  - Variables are added based on largest semi-partial correlation with the outcome and contribution to the model predictive power.

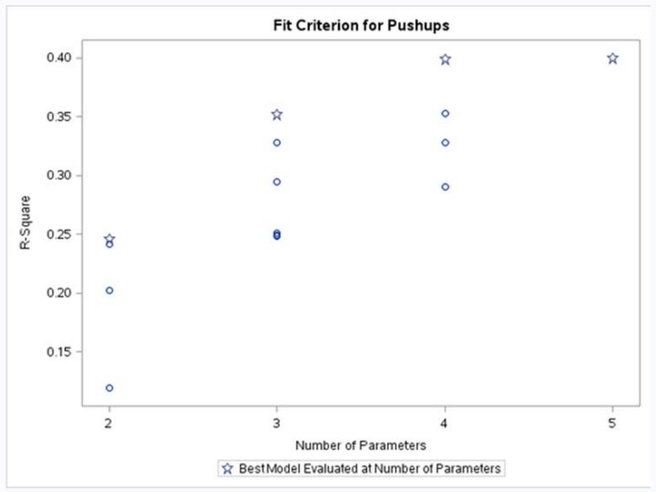
Model selection methods

- Backward method:
  - Starts with a model based on all predictors.
  - Variables are deleted based on a criterion linked to their significance.
  - Preferred to forward method.
- Stepwise method:
  - In SAS it is the same as the forward method, except the model is reassessed each time to see whether any redundant predictors can be removed.

Example: Model selection

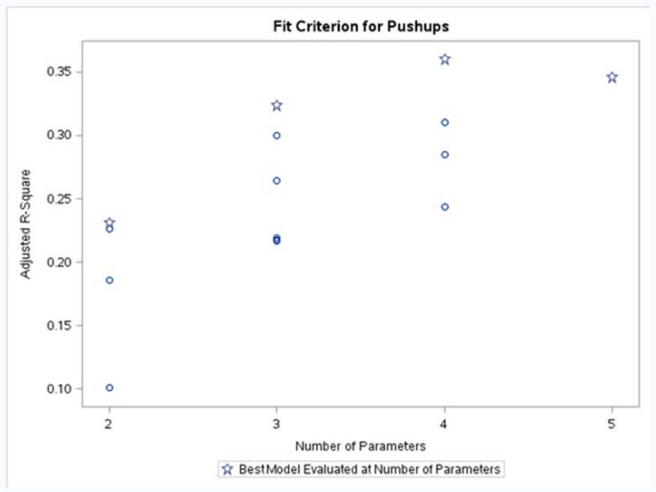
Model Index	Number in Model	R-Square	Adjusted R-Square	C(p)	Variables in Model
1	1	0.2464	0.2307	10.4785	Rest_Pulse
2	1	0.2420	0.2262	10.8103	Age
3	1	0.2026	0.1860	13.7615	Max_Pulse
4	1	0.1194	0.1011	19.9961	Run_Pulse
5	2	0.3517	0.3241	4.5862	Age Max_Pulse
6	2	0.3283	0.2997	6.3423	Age Rest_Pulse
7	2	0.2946	0.2646	8.8650	Age Run_Pulse
8	2	0.2510	0.2191	12.1369	Max_Pulse Rest_Pulse
9	2	0.2493	0.2174	12.2591	Max_Pulse Run_Pulse
10	2	0.2489	0.2169	12.2914	Rest_Pulse Run_Pulse
11	3	0.3995	0.3603	3.0068	Age Max_Pulse Run_Pulse
12	3	0.3527	0.3105	6.5096	Age Max_Pulse Rest_Pulse
13	3	0.3284	0.2846	8.3329	Age Rest_Pulse Run_Pulse
14	3	0.2901	0.2439	11.1998	Max_Pulse Rest_Pulse Run_Pulse
15	4	0.3996	0.3462	5.0000	Age Max_Pulse Rest_Pulse Run_Pulse

Example: Model selection



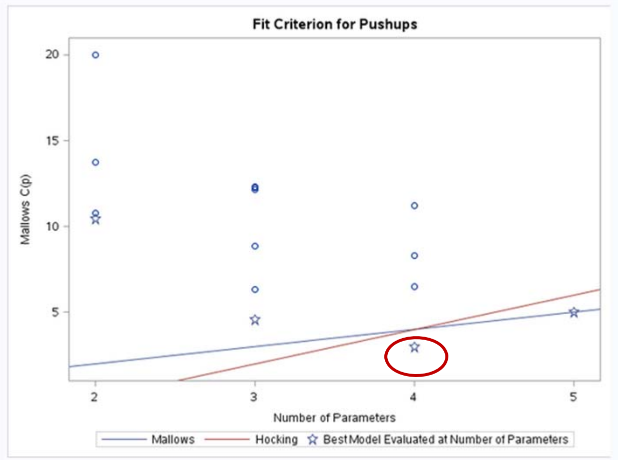
School of IT & Mathematical Sciences 35

Example: Model selection



School of IT & Mathematical Sciences 36

Example: Model selection



The best model according to both Mallow's and Hocking's criterion is a model with 4 parameters (3 predictors)

Multicollinearity

- Exists when there is a strong correlation between two or more predictors in a multiple regression model.
- The following problems arise:
  - Standard errors of regression coefficients increase.
  - Limited improvement in R^2.
  - It is difficult to assess the individual importance of a predictor.

Multicollinearity

- Variance inflation factor (VIF):
  - Indicates whether a predictor has a strong linear relationship with the other predictors.
  - Cause for concern: a value of 10 or higher.
- Tolerance statistic:
  - Reciprocal of VIF (or 1/VIF) .
  - Cause for concern: values below 0.1 indicate serious problems, values below 0.2 are worthy of concern.

Example: Full model

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	3175.89068	793.97267	7.49	0.0001
Error	45	4772.60932	106.05798		
Corrected Total	49	7948.50000			

Root MSE	10.29845	R-Square	0.3996
Dependent Mean	25.30000	Adj R-Sq	0.3462
Coeff Var	40.70532		

Example: Multicollinearity

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	93.56536	21.33252	4.39	<.0001	0
Age	1	-0.31264	0.10918	-2.86	0.0063	1.43496
Max_Pulse	1	-1.26238	0.54665	-2.31	0.0256	11.59795
Rest_Pulse	1	-0.02608	0.31565	-0.08	0.9345	4.34454
Run_Pulse	1	0.84388	0.45045	1.87	0.0675	8.19934

We have large VIF values for *Max_Pulse* and *Run_Pulse*.

We could either leave one of these variables out, or else create a new variable that is a linear combination of these two variables.

The extra sum of squares *F*-test

- Test the contribution of a specific set of variables by comparing the residuals of a full and a reduced model.

$$H_0 : \beta_{p+1} = \beta_{p+2} = \ldots = \beta_m = 0$$

$$H_1 : \text{at least one of the coefficients in not zero}$$

- The test statistic is

$$F = \frac{(SS_M^{full} - SS_M^{reduced})/(m-p)}{MS_R^{full}}$$

In our case, the comparison between the full model and Model 1 gives *F* = 7.17, which is statistically significant

p is the number of predictors in the reduced model, m is the number of predictors in the full model (excluding the intercept) and *n* is the number of observations

SAS code – regression with diagnostics

```
proc reg data=mydata.exercise
    plots(label only)=(cooksd studentbypredicted
                       dffits dfbetas);
    id Subj;
    model Pushups=Age Max_Pulse / CLB influence VIF;
    run;
quit;
```


SAS code – model selection

```
proc reg data=mydata.exercise
    plots(only)=(rsquare adjrsq cp);
    model Pushups=Age Max_Pulse Rest_Pulse Run_Pulse
            / selection=rsquare cp adjrsq;
    run;

proc reg data=mydata.exercise;
    Forward: model Pushups=Age Max_Pulse Rest_Pulse
      Run_Pulse / selection=forward;
    Backward: model Pushups=Age Max_Pulse Rest_Pulse
      Run_Pulse / selection=backward;
    Stepwise: model Pushups=Age Max_Pulse Rest_Pulse
      Run_Pulse / selection=stepwise;
    run;
quit;
```

Dummy coding

- For any categorical explanatory variable with *g* groups, only *g* – 1 terms should be included in the regression model:
  - Create g – 1 variables.
  - Choose one group as baseline, which is a group against which all other groups will be compared, so a control group or a group representing majority.
  - Assign the baseline group a value of 0 in all dummy variables.
  - For the first dummy variable, assign the first group the value of 1 and 0 for all the other groups. For the second dummy variable, assign the second group the value of 1 and 0 for all the other groups, and so on.

Example: Pulse rates

- Recall the pulse rates data set from Week 2.
- Suppose we wish to predict a person's pulse rate from their age and how often they exercise.
  - One predictor is numerical, the other categorical.
- Variable *Exercise* has three levels, coded 1 for 'high', 2 for 'moderate' and 3 for 'low'.
  - Make the low exercise group the baseline.
  - We need to create two dummy variables, which we will call *High* and *Moderate*.

SAS code: Dummy coding

```
/* Create dummy variable for level of exercise */

data work.pulse_rates_dummies;
    set mydata.pulse_rates;
    if exercise=1 then High=1;
    else High=0;
    if exercise=2 then Moderate=1;
    else  Moderate=0;
run;
```

SAS code: Dummy coding

```
/* List first 10 observations of the new data set */
proc print data=work.pulse_rates_dummies (obs=10)
      noobs;
    var Exercise High Moderate;
run;
```

Exercise	High	Moderate
2	0	1
2	0	1
1	1	0
1	1	0
3	0	0
3	0	0
2	0	1
2	0	1
1	1	0
2	0	1

To be sure the program worked correctly, PROC PRINT lists the first 10 observations

Example: Pulse rates

- Consider the following SAS output:

The REG Procedure
Model: MODEL1
Dependent Variable: Pulse

Number of Observations Read	110
Number of Observations Used	109
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1524.93702	508.31234	3.04	0.0324
Error	105	17572	167.35674		
Corrected Total	108	19097			

The model is statistically significant

Example: Pulse rates

- Consider the following SAS output:

Root MSE	12.93664	R-Square	0.0799
Dependent Mean	75.68807	Adj R-Sq	0.0536
Coeff Var	17.09205		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	90.66584	7.01683	12.92	<.0001
Age	1	-0.58565	0.31801	-1.84	0.0684
High	1	-10.56097	4.08552	-2.58	0.0111
Moderate	1	-2.96025	2.72670	-1.09	0.2801

Coefficients for *Age* and *High* are both statistically significant at 10% level.

Pulse = 90.67 – 0.59 Age – 10.56 High – 2.96 Moderate

Example: Pulse rates

Pulse = 90.67 – 0.59 Age – 10.56 High – 2.96 Moderate

- For the low exercise group, the relationship between pulse rate and age is:

  Pulse = 90.67 – 0.59 Age
- For the high exercise group, the pulse rate is on average 10.56 bpm lower than for the low exercise group.
- For the moderate exercise group, the pulse rate is on average 2.96 bpm lower than for the low exercise group.
  - This difference is however not statistically significant (P-value = 0.2801).

Multiple regression – some comments

- A great deal of care should be taken in selecting predictors for a model.
  - Values of regression coefficients depend on the variables in the model.
- Techniques used may depend upon the objectives of the analysis.
  - The focus when using iterative variable selection techniques is not the significance of each explanatory variable, but how well the overall model fits.
  - However, if the goal is to confirm a theory, other methods should be used.

Multiple regression – some comments

- Model selection decisions should never be left to a computer.
  - Models derived by a computer often take advantage of random sampling variation and there is also a danger of over-fitting as well as under-fitting.
- Diagnostic statistics should always be examined but it should be remembered that they are a way of assessing a model.
  - They should never be used to justify removing data points to achieve desirable change in regression parameters!

Multiple regression – some comments

- Checking assumptions is important if we want to generalise our regression model.
  - If assumptions have been violated, findings cannot be generalised beyond the sample.
  - It is still OK to use the model to draw conclusions about the sample even if assumptions are violated.