STAT 350

Assignment 6 Solutions

1. For the Nitrogen Output in Wallabies data set from Assignment 3 do forward, backward, stepwise and all subsets regression.

   Here is code for all the methods and with all subsets done both using C_p and using adjusted R^2.

```
data nit;
  infile 'nit.dat' ;
  input nitexc weight dryin wetin nitin ;
proc reg  data=nit;
   model nitexc = weight dryin wetin
         nitin /selection=FORWARD;
run ;
proc reg  data=nit;
   model nitexc = weight dryin wetin
         nitin /selection=BACKWARD;
run ;
proc reg  data=nit;
   model nitexc = weight dryin wetin
         nitin /selection=STEPWISE;
run ;
proc reg  data=nit;
   model nitexc = weight dryin wetin
         nitin /selection=CP;
run ;
proc reg  data=nit;
   model nitexc = weight dryin wetin
         nitin /selection=ADJRSQ;
run ;
```

   The conclusion of the output

```
          Forward Selection Procedure for Dependent Variable NITEXC

Step 1   Variable NITIN Entered     R-square = 0.95152988   C(p) =  0.19478831

                 DF         Sum of Squares      Mean Square          F   Prob>F

 Regression       1        176039.65105472  176039.65105472     451.52   0.0001
 Error           23          8967.30894528     389.88299762
 Total           24        185006.96000000
```

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	101.08658320	7.35602243	73626.68406474	188.84	0.0001
NITIN	0.64694573	0.03044597	176039.65105472	451.52	0.0001

Bounds on condition number: 1, 1

Step 2 Variable WETIN Entered R-square = 0.95383292 C(p) = 1.18772819

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	2	176465.72840308	88232.86420154	227.27	0.0001
Error	22	8541.23159692	388.23779986		
Total	24	185006.96000000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	88.74785286	13.87826796	15876.06942341	40.89	0.0001
WETIN	0.02934402	0.02801072	426.07734836	1.10	0.3062
NITIN	0.63199810	0.03356537	137640.57608325	354.53	0.0001

Bounds on condition number: 1.220562, 4.882248

No other variable met the 0.5000 significance level for entry into the model.

Summary of Forward Selection Procedure for Dependent Variable NITEXC

Step	Variable Entered	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	NITIN	1	0.9515	0.9515	0.1948	451.5192	0.0001
2	WETIN	2	0.0023	0.9538	1.1877	1.0975	0.3062

Backward Elimination Procedure for Dependent Variable NITEXC

Step 0 All Variables Entered R-square = 0.95426223 C(p) = 5.00000000

	DF	Sum of Squares	Mean Square	F	Prob>F

Regression	4	176545.15437710	44136.28859428	104.32	0.0001
Error	20	8461.80562290	423.09028114		
Total	24	185006.96000000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	111.02503203	92.09694447	614.87158016	1.45	0.2421
WEIGHT	-0.00594399	0.02807714	18.96195525	0.04	0.8345
DRYIN	-0.06000511	0.24052534	26.33224358	0.06	0.8055
WETIN	0.04783734	0.09083460	117.34499123	0.28	0.6042
NITIN	0.64809282	0.06471880	42427.40669891	100.28	0.0001

Bounds on condition number: 18.81062, 144.3604

Step 1 Variable WEIGHT Removed R-square = 0.95415974 C(p) = 3.04481775

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	176526.19242186	58842.06414062	145.70	0.0001
Error	21	8480.76757814	403.84607515		
Total	24	185006.96000000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	91.84916426	16.26623043	12876.34024832	31.88	0.0001
DRYIN	-0.08200363	0.21193002	60.46401878	0.15	0.7027
WETIN	0.05721277	0.07748286	220.18641282	0.55	0.4684
NITIN	0.65159457	0.06112979	45884.38908641	113.62	0.0001

Bounds on condition number: 15.29972, 84.5105

Step 2 Variable DRYIN Removed R-square = 0.95383292 C(p) = 1.18772819

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	2	176465.72840308	88232.86420154	227.27	0.0001
Error	22	8541.23159692	388.23779986		
Total	24	185006.96000000			

	Parameter	Standard	Type II

Variable	Estimate	Error	Sum of Squares	F	Prob>F
INTERCEP	88.74785286	13.87826796	15876.06942341	40.89	0.0001
WETIN	0.02934402	0.02801072	426.07734836	1.10	0.3062
NITIN	0.63199810	0.03356537	137640.57608325	354.53	0.0001

Bounds on condition number: 1.220562, 4.882248

Step 3 Variable WETIN Removed R-square = 0.95152988 C(p) = 0.19478831

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	1	176039.65105472	176039.65105472	451.52	0.0001
Error	23	8967.30894528	389.88299762		
Total	24	185006.96000000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	101.08658320	7.35602243	73626.68406474	188.84	0.0001
NITIN	0.64694573	0.03044597	176039.65105472	451.52	0.0001

Bounds on condition number: 1, 1

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination Procedure for Dependent Variable NITEXC

Step	Variable Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	WEIGHT	3	0.0001	0.9542	3.0448	0.0448	0.8345
2	DRYIN	2	0.0003	0.9538	1.1877	0.1497	0.7027
3	WETIN	1	0.0023	0.9515	0.1948	1.0975	0.3062

Stepwise Procedure for Dependent Variable NITEXC

Step 1 Variable NITIN Entered R-square = 0.95152988 C(p) = 0.19478831

	DF	Sum of Squares	Mean Square	F	Prob>F

Regression	1	176039.65105472	176039.65105472	451.52	0.0001
Error	23	8967.30894528	389.88299762		
Total	24	185006.96000000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	101.08658320	7.35602243	73626.68406474	188.84	0.0001
NITIN	0.64694573	0.03044597	176039.65105472	451.52	0.0001

Bounds on condition number: 1, 1

All variables left in the model are significant at the 0.1500 level.
No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Procedure for Dependent Variable NITEXC

Step	Variable Entered	Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	NITIN		1	0.9515	0.9515	0.1948	451.5192	0.0001

N = 25 Regression Models for Dependent Variable: NITEXC

C(p)	R-square	In	Variables in Model
0.19479	0.95152988	1	NITIN
1.18773	0.95383292	2	WETIN NITIN
1.56524	0.95296959	2	DRYIN NITIN
1.82386	0.95237815	2	WEIGHT NITIN
3.04482	0.95415974	3	DRYIN WETIN NITIN
3.06224	0.95411990	3	WEIGHT WETIN NITIN
3.27735	0.95362796	3	WEIGHT DRYIN NITIN
5.00000	0.95426223	4	WEIGHT DRYIN WETIN NITIN
103.27980	0.72493353	3	WEIGHT DRYIN WETIN
109.49540	0.70614534	2	DRYIN WETIN
183.76950	0.53171499	1	DRYIN
183.96969	0.53583097	2	WEIGHT DRYIN
314.82815	0.23657226	2	WEIGHT WETIN
324.50973	0.20985779	1	WETIN
414.13333	0.00489861	1	WEIGHT

```
-------------------------------------

 N = 25     Regression Models for Dependent Variable: NITEXC


   Adjusted     R-square        Variables in Model
   R-square                In

 0.94963591   0.95383292    2   WETIN NITIN
 0.94942249   0.95152988    1   NITIN
 0.94869409   0.95296959    2   DRYIN NITIN
 0.94804889   0.95237815    2   WEIGHT NITIN
 0.94761113   0.95415974    3   DRYIN WETIN NITIN
 0.94756560   0.95411990    3   WEIGHT WETIN NITIN
 0.94700338   0.95362796    3   WEIGHT DRYIN NITIN
 0.94511468   0.95426223    4   WEIGHT DRYIN WETIN NITIN
 0.68563831   0.72493353    3   WEIGHT DRYIN WETIN
 0.67943128   0.70614534    2   DRYIN WETIN
 0.51135477   0.53171499    1   DRYIN
 0.49363378   0.53583097    2   WEIGHT DRYIN
 0.17550378   0.20985779    1   WETIN
 0.16716974   0.23657226    2   WEIGHT WETIN
 -.03836666   0.00489861    1   WEIGHT
-------------------------------------
```

is that BACKWARD and STEPWISE settle for the model containing only Nitrogen Intake as a predictor. The forward selection method also includes Wet Intake because of the very high level of α (0.5) to enter. The all subsets method using C_p would settle on the model using only nitrogen intake but the adjusted R^2 method also includes Wet Intake. However, overall there seems little reason to include Wet Intake since it improves the fit very little and is not very significant at all.

2. Suppose X_1, X_2, X_3 are independent $N(\mu, \sigma^2)$ random variables, so that $X_i = \mu + \sigma Z_i$ with Z_1, Z_2, Z_3 independent standard normals.

   (a) If $X^T = (X_1, X_2, X_3)$ and $Z^T = (Z_1, Z_2, Z_3)$ express X in the form $AZ + b$ for a suitable matrix A and vector b.

   We have $A = \sigma I$ and $b^T = [\mu, \mu, \mu]$.

   (b) Show that X is $MVN_3(\mu_X, \Sigma_X)$ and identify μ_X and Σ_X.

   The definition of MVN is that X be of the form $AZ + b$ and then $\mu_X = b$ and $\Sigma_X = AA^T$. So $\mu_x^T = [\mu, \mu, \mu]$ and $\Sigma_X = \sigma^2 I$.

   (c) Let $Y_i = X_i - \bar{X}$ for $i = 1, 2, 3$ and $Y_4 = \bar{X}$. Show that $Y \sim MVN_4(\mu_Y, \Sigma_Y)$ and find μ_Y and Σ_Y.

The definition of MVN is that X be of the form $AZ+b$. Then if $Y = BX$ we have $Y = B(AZ + b) = (BA)Z + Bb$ so that Y is MVN with mean $\mu_Y = Bb = \mathrm{E}(Y)$ and $\Sigma_Y = (BA)(BA)^T = BAA^TB^T$. In this case we find

$$\mu_Y = \mathrm{E}(Y) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \mu \end{bmatrix}$$

and

$$B = \begin{bmatrix} 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

So

$$\Sigma_Y = \sigma^2 BB^T = \begin{bmatrix} 2/3 & -1/3 & -1/3 & 0 \\ -1/3 & 2/3 & -1/3 & 0 \\ -1/3 & -1/3 & 2/3 & 0 \\ 0 & 0 & 0 & 1/3 \end{bmatrix}.$$

(d) In class I may have stated that the if the covariance between two components of a multivariate normal vector is 0 then the components are independent, but I indicated a proof only when the multivariate normal distribution in question has a density. In this case the variance matrix is singular so there is no density. However, in terms of the original Z it is possible to find two independent functions of Z such that Y_1, Y_2, Y_3 are a function of the first function while Y_4 is a function of the second.

i. Let $U_1 = (Z_1 - Z_2)/\sqrt{2}$, $U_2 = (Z_1 + Z_2 - 2Z_3)/\sqrt{6}$ and $U_3 = (Z_1 + Z_2 + Z_3)/3$. Show that $U = (U_1, U_2, U_3)^T$ has a multivariate normal distribution and identify the mean and variance of U.
We have $U = AZ$ where

$$A = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 1/\sqrt{6} & 1/\sqrt{6} & -2/\sqrt{6} \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

Thus U is multivariate normal with mean 0 and variance covariance matrix AA^T. Multiply this out to check this is

$$\Sigma = AA^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/3 \end{bmatrix}$$

ii. Use the result in class, for multivariate normals which have a density to show that (U_1, U_2) is independent of U_3.
Since the variance covariance matrix is not singular we need only check that $\mathrm{Cov}(U_1, U_3) = \mathrm{Cov}(U_2, U_3) = 0$. These two entries in Σ are, indeed, 0.

iii. Express Y_3 as a function of U.
We have

$$Y_3 = X_3 - \bar{X} = \sigma(Z_3 - \bar{Z}).$$

Now

$$3U_3 - \sqrt{6}U_2 = 3Z_3$$

and

$$\bar{Z} = U_3.$$

Thus

$$Y_3 = \sigma(U_3 - \sqrt{6}U_2/3 - U_3) = -\sigma\sqrt{6}U_2/3.$$

iv. Use the fact that if X_1 and X_2 are independent then so are $G(X_1)$ and $H(X_2)$ for any functions G and H to show that Y_1, Y_2, Y_3 is independent of Y_4.
We see that Y_4 is a function of U_3. I claim that Y_1, Y_2 and Y_3 are each functions of U_1, U_2. You did Y_3 in the last part. Notice that

$$Y_2 = X_2 - \bar{X} = \sigma(Z_2 - \bar{Z})$$

and

$$Y_1 = X_1 - \bar{X} = \sigma(Z_1 - \bar{Z}).$$

Write

$$2U_3 + \sqrt{6}U_2/3 = Z_1 + Z_2$$

and then

$$2U_3 + \sqrt{6}U_2/3 + \sqrt{2}U_1 = 2Z_1$$

and

$$2U_3 + \sqrt{6}U_2/3 - \sqrt{2}U_1 = 2Z_2$$

These show

$$Z_1 - \bar{Z} = \sqrt{6}U_2/6 + \sqrt{2}U_1/2$$

and

$$Z_2 - \bar{Z} = \sqrt{6}U_2/6 - \sqrt{2}U_1/2$$

so both Y_1 and Y_2 are functions of U_1 and U_2.
Since Y_1, Y_2, Y_3 is a function of U_1, U_2 and Y_4 is a function of U_3 we see the desired independence.

v. Express the sample variance of the $X_i, i = 1, 2, 3$ in terms of U and use this to show that $(n-1)s_X^2/\sigma^2$ has a χ^2 distribution on 2 degrees of freedom (with $n = 3$). Note: in fact the sample variance of X_1, X_2 is a function of U_1. Generalizations of this idea can be used to develop an identity of the form

$$(n-1)s_n^2 = (n-2)s_{n-1}^2 + U_n^2$$

for a suitable U_n where s_n^2 is the sample variance for $X_1, \ldots, X_n$.

The sample variance is

$$s^2 = \frac{Y_1^2 + Y_2^2 + Y_3^2}{2}$$

Replace each Y_i by the formulas in the previous part to get that

$$\frac{2s^2}{\sigma^2} = (Z_1 - \bar{Z})^2 + (Z_2 - \bar{Z})^2 + (Z_3 - \bar{Z})^2$$

Write this in terms of U_1 and U_2 to get

$$\frac{2s^2}{\sigma^2} = \left\{\sqrt{6}U_2/6 + \sqrt{2}U_1/2\right\}^2 + \left\{\sqrt{6}U_2/6 - \sqrt{2}U_1/2\right\}^2 + \left\{\sqrt{6}U_2/3\right\}^2 = U_1^2 + U_2^2$$

This is a sum of squares of two independent normals each with mean 0 and variance 1 so we are done.

3. In class I discussed the general formula for a multivariate normal density. Suppose that Z_1 and Z_2 are independent standard normal variables. Assume that $X_1 = aZ_1 + bZ_2 + c$ and $X_2 = dZ_1 + eZ_2 + f$. Find the joint density of X_1 and X_2 by evaluating the formulas I gave in class. Express $P(X_1 \leq t)$ as a double integral. I want to see the integrand and the limits of integration but you need not try to do the integral.

The density in class was

$$\frac{1}{2\pi\sqrt{det(AA^T)}} \exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

where $\Sigma = AA^T$.

The entries in μ are c and f and we have

$$A = \begin{bmatrix} a & b \\ d & e \end{bmatrix}$$

Then

$$\Sigma = AA^T = \begin{bmatrix} a^2+b^2 & ad+be \\ ad+be & d^2+e^2 \end{bmatrix}$$

and

$$(AA^T)^{-1} = \begin{bmatrix} \frac{d^2+e^2}{(ae-bd)^2} & -\frac{ad+be}{(ae-bd)^2} \\ -\frac{ad+be}{(ae-bd)^2} & \frac{a^2+b^2}{(ae-bd)^2} \end{bmatrix}$$

Putting together all the algebra gives

$$f(x_1, x_2) = \frac{1}{2\pi|ae-bd|} \exp\left[-\frac{1}{2}\frac{q(x_1, x_2, a, b, c, d, e, f)}{(ae-bd)^2}\right].$$

where the exponent is the quadratic function

$$\begin{aligned} q(x_1, x_2, a, b, c, d, e, f) &= [(x_1 - c)^2(d^2 + e^2) \\ &\quad -2(ad+be)(x_1 - c)(x_2 - f) \\ &\quad +(x_2 - f)^2(a^2 + b^2)]. \end{aligned}$$

To compute $P(X_1 \leq t)$ we take the joint density of X_1 and X_2 and integrate it over the set of (x_1, x_2) such that $x_1 \leq t$ to get

$$P(X_1 \leq t) = \int_{-\infty}^{\infty} \int_{-\infty}^{t} f(x_1, x_2) dx_1 dx_2 \, .$$

4. Power and sample size calculations must be done before the data are gathered. However: pilot studies are often used to determine the size of the unknown parameters which are needed for these calculations. Use the Sand and Fibre Hardness data discussed in class as follows.

   (a) Consider the model

   $$Y_i = \beta_0 + \beta_1 S_i + \beta_2 F_i + \beta_3 F_i^2 + \epsilon_i$$

   Fit this model to get estimates of all the βs and of σ. Use these fitted values as if they were the true parameter values in the following.

   i. Compute the power of a two sided t test (at the 5% level) of the hypothesis that $\beta_3 = 0$ for

   $$\beta_3 \in \{-0.006, -0.003, 0, 0.003, 0.006\}.$$

   We need the non-centrality parameter

   $$\delta = \frac{\beta_3}{\sigma \sqrt{a^T (X^T X)^{-1} a}}$$

   where $a^T = [0\ 0\ 0\ 1]$. We get β_3 values from the question and the denominator from the standard error of $\hat{\beta}_3$. I find that standard error to be 0.001995. This gives δ values equal to -3,-1.5,0,1.5,3 (close enough).
   Turning to Table B.5 with 14 degrees of freedom for error I get powers 0.8, somewhere between 0.15 and 0.46, 0.05, somewhere between 0.15 and 0.46 and 0.8. The "somewheres" are, I guess in the range of 0.3-0.35.

   ii. Find a number m of copies of the basic design (each combination of Sand and Fibre tried twice) to guarantee that the power of the t-test of the hypothesis $\beta_3 = 0$ is 0.9 when the true parameter values are as in the fit above.
   The fitted value of β_3 is -0.003733 and the corresponding value of δ is just the t statistic in the computer output for testing $\beta_3 = 0$. This is -1.871 though we use 1.871 in the tables because our tests are to be two sided. We need $1.871\sqrt{m}$ to give us a power of 0.9. For large numbers of degrees of freedom for error this power is about half way between the figure under $\delta = 4$ and the figure under $\delta = 3$ so we solve $1.871\sqrt{m} = 3.5$ and get $m = 3.5$. We would have to use $m = 4$ giving $\delta = 3.74$ and a power closer to 0.95. However we could treat the basic design as having 9 points and try 63 points in total (7 at each combination of S and F) and get a power quite close to 0.9.

(b) Now consider the model

$$Y_i = \beta_0 + \beta_1 S_i + \beta_2 F_i + \beta_3 F_i^2 + \beta_4 S_i^2 + \beta_5 S_i F_i + \epsilon_i$$

Fit this model to get estimates of all the βs and of σ. Use these fitted values as if they were the true parameter values in the following.

Find a number m of copies of the basic design (each combination of Sand and Fibre tried twice) to guarantee that the power of the F-test of the hypothesis $\beta_3 = \beta_4 = \beta_5 = 0$ is 0.9 when the true parameter values are as in this fit.

We can use table B.11 on page 1338 since we will have 3 numerator degrees of freedom for this F test. We will have $18m - 6$ *degrees of freedom for error and so will need a* ϕ*, in the notation of the table, quite close to 2. This makes* $\delta^2 = (p+1)2^2 = 16$*. The estimates are* $\hat{\beta}_3 = -0.003733$*,* $\hat{\beta}_4 = -0.004815$ *and* $\hat{\beta}_5 = -0.001000$*. To get our value of* δ^2 *for the basic design I regress* $\hat{\beta}_3 F^2 + \hat{\beta}_4 S^2 + \hat{\beta}_5 SF$ *on* S *and* F *and find the error sum of squares is 27.58. Divide this by* $\hat{\sigma}^2 = 6.77$ *to find*

$$\delta^2 = 4.07$$

.

I need to get up to 16 so I need $m = 4$*.*