

ST512

Summer Session II, 2008

Quiz 2

Name: _____

Directions: Answer questions as directed. Please show work. Give expressions for answers where possible, as partial credit may be awarded in cases where expressions are correct, but numerical answers are not.

You may use the back of the page if you need extra space.

1. An NCSU entomologist selects $N = 30$ homes from a large population of local houses that are similar in age and occupancy, then randomly assigns them to three treatment groups (below) and measures the reductions in trap counts over a two week period.

Group	Symbol	Extermination strategy
1	C	control (no treatment)
2	KB	poison applied in kitchen and bathroom
3	W	poison applied in whole house (same amount as KB)

Let y_{ij} denote the observed reduction for house j receiving strategy i . Then

sums of squares	sample means and variances
$\sum_{i=1}^3 \sum_{j=1}^{10} (\bar{y}_{i+} - \bar{y}_{++})^2 = 71.5$	$\bar{y}_{1+} = 1.85$, $s_1^2 = 6.86$
$\sum_{i=1}^3 \sum_{j=1}^{10} (y_{ij} - \bar{y}_{++})^2 = 216.0$	$\bar{y}_{2+} = 2.98$, $s_2^2 = 7.40$
$\sum_{i=1}^3 \sum_{j=1}^{10} (y_{ij} - \bar{y}_{i+})^2 = 144.5$	$\bar{y}_{3+} = 5.54$, $s_3^2 = 1.82$

- (a) Compose an analysis of variance (ANOVA) table with four columns: source of variation, degrees of freedom, sum of squares, and mean square. Test the hypothesis that all three strategies lead to the same average reduction in roach counts. Report a p -value, using an F -table, or an applet, or software. Draw a brief conclusion.

Source	df	Sum of squares	Mean Square
Treatment	2	71.5	35.75
Error	27	144.5	5.35
Total	29	216	

In testing $H_0 : \mu_C = \mu_{KB} = \mu_W$, we observe $F = 35.75/5.35 = 6.68$ with a p -value of .0044 on $df = 2, 27$. Therefore, the 3 sample treatment means differ significantly. That is, there is evidence of a treatment effect is highly significant.

- (b) Consider a model for the mean reduction in roach counts that uses multiple linear regression with two indicator variables, one for the conventional strategy, X_{KB} and another for the whole house strategy, X_W . This model was fit using PROC REG and output for the regression coefficients is included on the next page.

- i. Estimate three mean differences: between C and W , between C and KB and between W and KB .

Difference	Estimate	\widehat{SE}
$W - C$	3.69	1.03
$KB - C$	1.13	1.03
$W - KB$	$3.69 - 1.13 = 2.56$	1.03

- ii. Standard error may also be obtained from

$$\widehat{SE}(\bar{y}_{i+} - \bar{y}_{j+}) = \sqrt{2MS[E]/10}$$

or

$$\widehat{SE}(\hat{\beta}_W - \hat{\beta}_{KB}) = \sqrt{0, 1, -1) \hat{\Sigma} \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}}$$

2. The relationship between fuel efficiency (**mpg**) and $p = 10$ automotive characteristics is modelled using multiple linear regression with a survey of $n = 32$ different cars.

Variable	meaning	Variable	meaning
wt	Weight (<i>lbs/1000</i>)	raxratio	Rear axel ratio
qmsec	1/4 mile time	cyl	Number of cylinders
standard	Trans. (0=auto, 1>manual)	straight	engine (0=v, 1=straight)
disp	Displacement (cu.in.)	gears	Number of forward gears
hp	Gross horsepower	carbs	Number of carburetors

A multiple linear regression of **mpg** on all $p = 10$ explanatory variables is fit (last page):

$$\begin{aligned} & \mu(wt, qmsec, standard, disp, hp, raxratio, cyl, straight, gears, carbs) \\ &= \beta_0 + \beta_1 wt + \beta_2 qmsec + \beta_3 standard + \beta_4 disp + \beta_5 hp \\ &+ \beta_6 raxratio + \beta_7 cyl + \beta_8 straight + \beta_9 gears + \beta_{10} carbs \end{aligned}$$

In terms of matrices, the model may be written $Y = X'\beta + E$ where X is a design matrix and E is a vector of independent normal errors with constant variance, σ^2 .

- (a) Specify the dimension of each matrix: Y, X and $(X'X)^{-1}$.

$$dim(Y) = (32 \times 1), \quad dim(X) = (32 \times 11), \quad dim((X'X)^{-1}) = (11 \times 11)$$

- (b) Report the transpose of the vector $(X'X)^{-1}X'Y$.

$$\hat{\beta}' = (12.3, -3.7, 0.8, 2.5, .01, -.02, .79, -.11, .32, .66, -.2)$$

- (c) Are any of the partial regression coefficients significant? Pick one of them, report the p -value and interpret it. (Explain what hypothesis is being tested by this p -value and clarify what models are being compared. Use level $\alpha = .05$) The p -value for a test of the partial regression coefficient for weight, (**wt**), $\beta_1 = 0$ is not less than .05, so there is no evidence that mean fuel efficiency depends on weight, after controlling for the other nine predictors.
- (d) Consider the hypothesis where efficiency does not depend on any of these $p = 10$ characteristics. Give the null hypothesis (H_0) in terms of regression coefficients. Report the F -ratio and p -value for a test of this hypothesis.

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_{10} = 0$$

$$F = 13.9, p < .0001$$

- (e) Consider a reduced model with only three predictors: **wt** and **qmsec** and **standard**.
- Conduct a statistical test to compare this model with the full model. In doing so, specify the null hypothesis clearly. Say what level of significance you choose. Draw a conclusion regarding the comparison.

$$H_0 : \beta_4 = \beta_5 = \dots = \beta_{10} = 0$$

$$F = \frac{(SS[R]_f - SS[R]_r)/(10 - 3)}{MS[E]_f} = \frac{(978.6 - 956.8)/7}{7.02} = 0.44(df = 7, 21)$$

After controlling for **wt** and **qmsec** and **standard**, there is no evidence of dependence of mean efficiency on any of the other predictors.

- Using the output entitled "MODEL2", test the hypothesis that efficiency does not depend on any of the **3** characteristics in the reduced model. Report an F -ratio and p -value and draw a brief conclusion.

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

$$F = 52.8, p < .0001, df = 3, 28$$

Therefore, efficiency is related to at least one of the three predictors.

- Using the reduced model, estimate the mean fuel efficiency among cars that weigh *2000lbs*, take 20 seconds for the quarter mile and have an automatic transmission.

$$\begin{aligned} \hat{\mu}(\text{wt} = 2, \text{qmsec} = 20, \text{standard} = 0) &= \hat{\beta}_0 + \hat{\beta}_1(2) + \hat{\beta}_2(20) \\ &= 9.62 - 3.92(2) + 1.23(20) + 2.94(0) \\ &= 26.4\text{mpg} \end{aligned}$$

- Provide a matrix expression that may be evaluated to obtain a standard error for your answer to part (e), **iii**.

$$\widehat{SE}(\hat{\mu}) = \sqrt{MS[E](1, 2, 20, 0)(X'X)^{-1} \begin{pmatrix} 1 \\ 2 \\ 20 \\ 0 \end{pmatrix}}$$

- Fill in the two missing elements in the estimated variance-covariance matrix of the regression coefficients, $\hat{\Sigma}$ given in the output (AAAA and BBBB). $AAAA = (1.41)^2$, $BBBB = -6.867$. (First is $SE(\hat{\beta}_2)^2$, second is by symmetry.
- Think about a plot of the observed efficiencies against the predicted values. What is the squared correlation from such a plot? $r^2 = 957/1126 = .85$
- It seems the reduced model enables us to detect some important explanatory variables but the full model does not. What is the problem with the full model? Use between 1 and 20 words in your answer. **The full model is overfit and there is multicollinearity.**

```

proc reg ;
  model mpg=wt qmsec standard disp hp raxratio cyl straight gears carbs;
  model mpg=wt qmsec standard/covb;
run;

```

The REG Procedure
Model: MODEL1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	978.55276	97.85528	13.93	<.0001
Error	21	147.49443	7.02354		
Corrected Total	31	1126.04719			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	12.30337	18.71788	0.66	0.5181
wt	1	-3.71530	1.89441	-1.96	0.0633
qmsec	1	0.82104	0.73084	1.12	0.2739
standard	1	2.52023	2.05665	1.23	0.2340
disp	1	0.01334	0.01786	0.75	0.4635
hp	1	-0.02148	0.02177	-0.99	0.3350
raxratio	1	0.78711	1.63537	0.48	0.6353
cyl	1	-0.11144	1.04502	-0.11	0.9161
straight	1	0.31776	2.10451	0.15	0.8814
gears	1	0.65541	1.49326	0.44	0.6652
carbs	1	-0.19942	0.82875	-0.24	0.8122

Model: MODEL2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	956.76126	318.92042	52.75	<.0001
Error	28	169.28593	6.04593		
Corrected Total	31	1126.04719			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	9.61778	6.95959	1.38	0.1779
wt	1	-3.91650	0.71120	-5.51	<.0001
qmsec	1	1.22589	0.28867	4.25	0.0002
standard	1	2.93584	1.41090	2.08	0.0467

Covariance of Estimates

Variable	Intercept	wt	qmsec	standard
Intercept	48.435934514	-3.681623712	-1.8831754	-6.867614794
wt	-3.681623712	0.5058077651	0.0976336178	0.7672015854
qmsec	-1.8831754	0.0976336178	0.0833301114	0.2011700148
standard	BBBBBBB	0.7672015854	0.2011700148	AAAAAAA