

1. (Rao example 11.15) A study measures oxygen demand (y) (on a log scale) and five explanatory variables (see below). Data are available as

x_1 : biol. O_2 demand
 x_2 : total Kjeldahl Nitrogen
 x_3 : total solids
 x_4 : total volatile solids
 x_5 : chem. O_2 demand

- (a) Among x_1 through x_5 , which variables exhibit a significant linear association with y ?
- (b) Fit a “full” multiple linear regression model which includes all of the variables you identified in part (a). For each partial regression coefficient, report the p -value for a test that the coefficient is 0. (The problem is even worse if you fit the model with all five of $x_1 - x_5$. Watch what happens to these partial slope p -values in this case.)
- (c) Use an F -test to compare the nested (reduced) model $\mu = \beta_0 + \beta_2x_2 + \beta_4x_4$ with the “full” model **with all five predictors**. Is there enough evidence against $\beta_1 = \beta_3 = \beta_5 = 0$ to reject the reduced model?
- (d) Consider the model involving all five predictors, x_1, x_2, x_3, x_4, x_5 . How many subsets with at least one predictor are possible? (Answer: a lot!)
- (e) Use the C_p criterion (or any other reasonable model selection criteria) to choose the best subset model for predicting log-oxygen demand:

```
proc reg;  
  model y=x1-x5/selection=cp;  
run;
```

- (f) Consider the model $\mu = \beta_0 + \beta_3x_3 + \beta_5x_5$.
- Estimate the mean log-oxygen demand when $x_3 = 5$ and $x_5 = 6$.
 - Report a standard error. Give the product of vectors and matrices that is evaluated to get this standard error.
 - Estimate the standard deviation of log-oxygen demand for $x_3 = 5$ and $x_5 = 6$.
 - Obtain a plot of the residuals against the fitted values.
 - Obtain a normal plot of the residuals (use the `normal` option in `proc univariate`.)
- (g) Estimate the difference between the slope for x_3 and that for x_5 .
- (h) Fit a simple linear regression model with $x_3 + x_5$ as the single predictor. Estimate the standard deviation of log-oxygen demands when $x_3 + x_5 = 11$. Compare with earlier question.
- (i) **optional:** For a similar challenge, see the NFL problem on the course website (and compare the predicted outcome for the Colts-Patriots game with the observed outcome).

2. Rao 12.5: (Refer to “plantht1.dat” and Example 8.2 in Rao) Four randomly selected seedlings were grown under $t = 5$ experimental conditions and heights at four weeks were measured:

t	Label	Description	Sample mean	Sample variance
1	D	Darkness	34.02	2.73
2	AL	safelight type A, low intensity	31.9	0.87
3	AH	safelight type A, high intensity	30.84	0.15
4	BL	safelight type B, low intensity	34.31	0.20
5	BH	safelight type B, high intensity	34.29	1.52

- Write a general linear model using dummy variables.
 - Write a general linear model using factorial effects.
 - Conduct an F -test for the null hypothesis that none of the treatments have any effect on mean plant height.
 - Among the non-darkness treatments, express the mean difference between low and high intensity as a function of parameters in part (a) above. Also, do this for part (b). Report an estimate of this effect, along with a standard error.
3. Consider the sample of $n_F = 18$ girls and $n_M = 22$ boys in “Bigclass.txt” as a random sample from a population of interest.
- Use regression with an indicator variable to conduct an equal variances t -test of the hypothesis the average heights of the two populations (boys and girls) are equal. Is this hypothesis plausible in light of these data? Also, do this with software for a two-sample comparisons of means like PROC TTEST and compare the results. How is the pooled sample variance from the two-sample comparison of means related to the error mean square from the regression?
 - Conduct a linear regression of height on age, ignoring gender. Is there a significant linear association between height and age? Report a p -value.
 - Fit a model to test the hypothesis of equal mean heights for boys and girls of the same age when assuming the same dependence of height on age for boys as for girls
 - Test that this dependence on age is constant across the two genders.

4. Rao 12.3a A sample of $n = 30$ subjects were randomly assigned to three therapies/treatments for improving mental capacity. On each subject, pretest (Z) and posttest (Y) measurements were made. The data are available on the ST512 website as “raoeg12.1.dat.” The code below should suffice for reading them in:

```
data scores;
  do therapy=1 to 3; /* loops over three pairs of columns */
    input z y @; /* @ prevents datastep from reading new record */
    output; /* write current values of variables to datastep */
  end;
run;
```

- (12.3a) Construct the one-way ANOVA table for comparing the three treatment means when z is ignored.

```
proc glm;
  class therapy;
  model y=therapy;
  means therapy; /* will add treatment means to the output */
run;
```

- (a) Conduct an F -test for equality of means. That is, specify a model and a null hypothesis for no therapy effect, then compute the F -ratio

$$F = MS(trt)/MS(E), \quad df = 2, 27$$

and compare it to the critical value $F(.05, 2, 27)$.

- (b) Consider the following equivalent model, that leads to the same inference regarding treatment effects,
- (c) Plot y versus z with a different symbol for each therapy.
- (d) Using PROC GLM or PROC REG, fit the following analysis of covariance (ANCOVA) model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 z_i + E_i$$

where X_{i1} and x_{i2} are indicator variables for therapies 1 and 2 respectively. Report each regression coefficient along with a standard error.

- (e) Report the F -test for a therapy effect, after controlling for the effect of the pretreatment (z) score.
- (f) Report the unadjusted post-test score for therapy 2.
- (g) Report the adjusted post-test score for therapy 2, along with a standard error.