

Statistics 512 more review problems for the final exam, Solution

1. Short answer questions. Each part is unrelated.

- (a) The true means (μ_1, μ_2, μ_3) in a cell means model have values 40, 50, and 40, respectively. The error variance σ^2 is 20, and the design is balanced with 4 replicates per level. What is the distribution of the following contrast estimate?

$$\hat{L} = \bar{Y}_1 + \bar{Y}_2 - 2\bar{Y}_3$$

Solution: The estimate \hat{L} is **normal** with mean $40 + 50 - 2 \times 40 = 10$ and variance

$$\begin{aligned} \sigma^2 \sum \frac{c_i^2}{n} &= 20 \left(\frac{1}{4} + \frac{1}{4} + \frac{2^2}{4} \right) \\ &= (20) \left(\frac{6}{4} \right) \\ &= \mathbf{30}. \end{aligned}$$

- (b) In a 2×2 ANOVA, the population means are $\mu_{1,1} = 20$, $\mu_{2,1} = 25$, $\mu_{1,2} = 31$, and $\mu_{2,2} = 36$. Is there an interaction in this model? Explain why or why not.

Solution: No, there is no interaction, since each interaction effect is 0.

$$\begin{aligned} \mu &= \frac{20+25+31+36}{4} = 28 \\ \mu_{1.} &= 25.5 & \mu_{2.} &= 30.5 \\ \mu_{.1} &= 22.5 & \mu_{.2} &= 33.5 \\ \alpha\beta_{1,1} &= 20 - 25.5 - 22.5 + 28 = 0 \\ \alpha\beta_{2,1} &= 25 - 30.5 - 22.5 + 28 = 0 \\ \alpha\beta_{1,2} &= 31 - 25.5 - 33.5 + 28 = 0 \\ \alpha\beta_{2,2} &= 36 - 30.5 - 33.5 + 28 = 0 \end{aligned}$$

- (c) After performing a least squares regression, you find that the squares of the residuals (r_i^2) are linearly related to the predictor X according to the following equation:

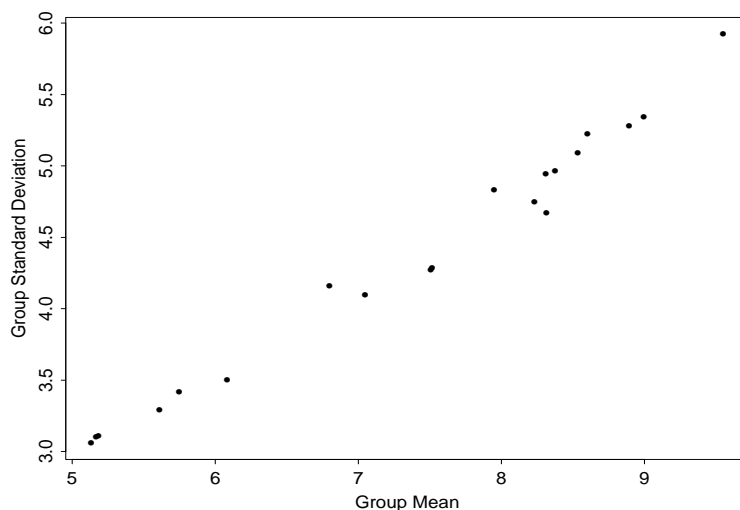
$$E(r_i^2) \approx 1 + X_i.$$

You decide to use weighted linear regression in relating the response Y to the predictor X . Give a formula for the weights w_i in terms of the variables in the model. (Assume all the X_i are positive.)

Solution: Since $E(r_i^2) = \sigma_i^2$, we use the fact the weights should be approximately proportional to $1/\sigma_i^2$ or

$$\frac{1}{r^2} = \frac{\mathbf{1}}{\mathbf{1} + \mathbf{X}_i}.$$

- (d) Given the following plot of μ_i versus σ_i , what transformation of the response would you recommend to make the variance constant?



Solution: Since it appears that μ_i is proportional to σ_i , the response should be **log-transformed** to make the variance constant.

- (e) A two-way **additive** ANOVA model has 3 and 4 levels for the two factors. If the data have 25 observations (i.e., $n_T = 25$), what is the error degrees of freedom?

Solution: The model degrees of freedom is $(3 - 1) + (4 - 1) = 5$. The total degrees of freedom is $25 - 1 = 24$. The error degrees of freedom is thus $24 - 5 = \mathbf{19}$.

- (f) Show that the criterion $C_p \leq p$ is equivalent to the criterion

$$\frac{MSE_p}{MSE(Full)} \leq 1.$$

Solution: Using the fact that $MSE_p = \frac{SSE_p}{n-p}$, it follows that $C_p \leq p$ implies that

$$\begin{aligned} \frac{SSE_p}{MSE(Full)} - (n - 2p) &\leq p \text{ (definition of } C_p) \\ \frac{SSE_p}{MSE(Full)} &\leq n - p \\ \frac{SSE_p/(n - p)}{MSE(Full)} &\leq 1 \\ \frac{MSE_p}{MSE(Full)} &\leq 1 \end{aligned}$$

- (g) The price per unit increases with lot size until the size is 200. The price per unit stays constant for lot sizes greater than 200. (In other words, the function relating

price to lot size is a piecewise linear function. The slope of the line for lot size greater than 200 is zero.)

The output from `proc print` after reading in the data is

Obs	cost	lotsize	cslope
1	128.7	100	
2	107.7	125	
3	85.0	150	
4	70.3	175	
5	48.3	200	
6	46.0	225	
7	47.6	250	
8	47.9	275	
9	48.3	300	

Fill in an additional column of predictor values (under the `cslope` heading) that you would use to fit the previously described piecewise linear model. What `model` statement would you use in `proc reg` for this model?

Solution: The column values under `cslope` are

cslope
100
125
150
175
200
200
200
200
200
200

In this case, the `model` statement would be `model cost = cslope.`

2. Refer to the SAS output marked OUTPUT FOR PROBLEM 2.

- (a) Write the factor effects model used for this analysis. Include numerical values for the number of levels being compared and the numbers of observations per treatment.

Solution: The factor effects model is

$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + \alpha\beta_{i,j} + \epsilon_{i,j,k},$$

where each independent $\epsilon_{i,j,k} \sim N(0, \sigma^2)$ for $i = 1, 2, 3$, $j = 1, 2$, $k = 1, \dots, 4$.

- (b) Summarize the results of the hypothesis tests for main effects and interactions.

Solution: The main effects A and B are both significant ($p = 0.0363$ and $p < 0.0001$, respectively), but the interaction effect $A \times B$ is not significant ($p = 0.5665$).

(c) Explain the results of the Tukey procedures for the main effects.

Solution: For factor A , levels 1 and 2 are significantly different from one another, but neither is significantly different from level 3.

For factor B , the two levels are significantly different from one another.

(d) Estimate the residual variance σ^2 .

Solution: The estimate of residual variance is $\hat{\sigma}^2 = MSE = \mathbf{3161.8997}$.

3. Refer to the SAS output marked OUTPUT FOR PROBLEM 3. Greenhouse benches were set up as blocks. Within each block, plants of different genetic varieties were grown. The maximum height of each plant was measured.

(a) Write the cell means model for this analysis. Include numerical values for the number of levels being compared and the numbers of observations per treatment. Also state the distributional assumption.

Solution: The cell means model is

$$Y_{i,j,k} = \mu_{i,j} + \epsilon_{i,j,k}$$

where the $\epsilon_{i,j,k}$ are independent with distribution $N(0, \sigma^2)$ for $i = 1, \dots, 4$, $j = 1, \dots, 6$, and $k = 1$.

(b) Explain why no interaction term was included in the model. Describe graphical evidence that would justify this assumption (of no interaction)?

Solution: With only one observation per treatment, the interaction effect cannot be estimated due to insufficient degrees of freedom. (An interaction effect would take up $(4 - 1)(6 - 1) = 15$ degrees of freedom. In such a model, the error degrees of freedom would become 0.)

To justify the assumption of no interaction, we would need to see an interaction plot that shows the cell means (which are made up of individual values) vary in parallel with differing levels of each factor.

(c) Write the factor effects model used for this analysis. Then, give a numerical estimate for each parameter in the model.

Solution: The factor effects model used for this analysis is

$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + \epsilon_{i,j,k},$$

where $\epsilon_{i,j,k} \sim N(0, \sigma^2)$.

The parameter estimates are

$$\begin{aligned}\hat{\sigma}^2 &= \mathbf{0.424} & \hat{\beta}_1 &= 18.2 - 17.1 = \mathbf{1.1} \\ \hat{\mu} &= \mathbf{17.1} & \hat{\beta}_2 &= 16.4 - 17.1 = \mathbf{-0.7} \\ \hat{\alpha}_1 &= 18.0 - 17.1 = \mathbf{0.9} & \hat{\beta}_3 &= 16.3 - 17.1 = \mathbf{-0.8} \\ \hat{\alpha}_2 &= 21.0 - 17.1 = \mathbf{3.9} & \hat{\beta}_4 &= 17.15 - 17.1 = \mathbf{0.05} \\ \hat{\alpha}_3 &= 15.5 - 17.1 = \mathbf{-1.6} & \hat{\beta}_5 &= 18.6 - 17.1 = \mathbf{1.5} \\ \hat{\alpha}_4 &= 13.9 - 17.1 = \mathbf{-3.2} & \hat{\beta}_6 &= 15.95 - 17.1 = \mathbf{-1.15}.\end{aligned}$$