**STAT 741** **MIDTERM SOLUTIONS**

1. Consider the linear model $Y_{ij} = \beta_0 + \beta_1 x_i + \varepsilon_{ij}$ for $i = 0, 1$, $j = 1, \ldots, n$, where $x_0 = 0$ and $x_1 = 1$. Assume that the $\varepsilon_{ij}$ are i.i.d. $N(0, \sigma^2)$.

(a) Find the least squares estimators of $\beta_0$ and $\beta_1$ and an unbiased estimator of $\sigma^2$.

(b) Show that the test of $H_0$: $\beta_1 = 0$ is the usual two sample Student $t$ test for the equality of means.

**Solution.** (a)

$$\begin{aligned}\mathcal{S} &= \sum_{i=0}^{1}\sum_{j=1}^{n}(Y_{ij} - \beta_0 - \beta_1 x_i)^2 = \sum_{i=0}^{1}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i\cdot} + \bar{Y}_{i\cdot} - \beta_0 - \beta_1 x_i)^2 \\ &= \sum_{i=0}^{1}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i\cdot})^2 + n(\bar{Y}_{0\cdot} - \beta_0)^2 + n(\bar{Y}_{1\cdot} - \beta_0 - \beta_1)^2\end{aligned}$$

Here we use the fact that $x_i$ is either zero or one. Obviously $\mathcal{S}$ is minimized if $\beta_0 = \bar{Y}_{0\cdot}$ and $\beta_0 + \beta_1 = \bar{Y}_{1\cdot}$. Then the least squares estimates are $\hat{\beta}_0 = \bar{Y}_{0\cdot}$ and $\hat{\beta}_1 = \bar{Y}_{1\cdot} - \bar{Y}_{0\cdot}$. The estimator of the variance is $s^2 = \sum\sum(Y_{ij} - \bar{Y}_{i\cdot})^2/(2n-2)$.

In matrix terms, let $\mathbf{1}_n$ denote an $n$-dimensional vector with all components equal to 1. Then

$$\mathbf{X} = \begin{bmatrix} \mathbf{1}_n & \mathbf{0}_n \\ \mathbf{1}_n & \mathbf{1}_n \end{bmatrix}; \quad \mathbf{X}^T\mathbf{X} = \begin{bmatrix} 2n & n \\ n & n \end{bmatrix}; \quad \mathbf{X}^T\mathbf{Y} = \begin{bmatrix} \sum_{i=0}^{1}\sum_{j=1}^{n} Y_{ij} \\ \sum_{j=1}^{n} Y_{1j} \end{bmatrix}.$$

Then the estimates are given by $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.

(b) The test can be constructed in various ways. Note that $\operatorname{Var}\hat{\beta}_1 = 2\sigma^2/n$ and that $s^2$ is the usual pooled variance in the two sample Student $t$ test. Therefore the test statistic based on the confidence interval method is

$$t = (\bar{Y}_{1\cdot} - \bar{Y}_{0\cdot})/(s\sqrt{2/n}).$$

The null model is $Y_{ij} = \beta_0 + \varepsilon_{ij}$ and the least squares estimate is $\bar{Y}_{\cdot\cdot}$. The numerator of the $F$ statistic is $n(\bar{Y}_{0\cdot} - \bar{Y}_{\cdot\cdot})^2 + n(\bar{Y}_{1\cdot} - \bar{Y}_{\cdot\cdot})^2$ with 1 degree of freedom. After some algebra, this reduces to $n(\bar{Y}_{1\cdot} - \bar{Y}_{0\cdot})^2/2$ so that $\mathcal{F} = t^2$.

2. Suppose that one considers a two factor additive model $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$, $i = 1, 2, 3$, $j = 1, 2, 3$, but that not all combinations $(i, j)$ of Factors $A$ and $B$ are observed. In each of the following cases decide whether or not all contrasts in the $\alpha_i$ and $\beta_j$ are estimable. That is, decide on the estimability of all parametric functions of the form $\alpha_i - \alpha_{i'}$ and $\beta_j - \beta_{j'}$.

Case I

| | | |
|---|---|---|
| x | x | – |
| x | x | x |
| – | x | x |

Case II

| | | |
|---|---|---|
| x | x | – |
| – | x | x |
| – | – | x |

Case III

| | | |
|---|---|---|
| x | x | – |
| x | x | – |
| – | – | x |

In each case the symbol x indicates an observed factor combination, and the symbol – indicates an unobserved factor combination.

**Solution.** First note that $E[Y_{ij} - Y_{ij'}] = \beta_j - \beta_{j'}$. Therefore in Case I all differences $\beta_j - \beta_{j'}$ are estimable because for some $i$, $Y_{i,j}$ and $Y_{i,j'}$ are both observed. A similar argument applies to $\alpha_i - \alpha_{i'}$.

In Case II, $\alpha_1 - \alpha_2$ is estimated by $Y_{12} - Y_{22}$ and $\alpha_2 - \alpha_3$ is estimated by $Y_{23} - Y_{33}$. Then one can estimate $\alpha_1 - \alpha_3$ by $(Y_{12} - Y_{22}) + (Y_{23} - Y_{33})$. Similarly all differences of $\beta$'s are estimable.

In Case III, $\alpha_1 - \alpha_2$ and $\beta_1 - \beta_2$ can be estimated as in Case I. To see that $\alpha_1 - \alpha_3$ is not estimable, suppose to the contrary that $E[\sum_i \sum_j c_{ij} Y_{ij}] = \alpha_1 - \alpha_3$. Then the coefficient of $\beta_3$ must be zero in this expectation. But $\beta_3$ appears only in $E[Y_{33}]$, so we must have $c_{33} = 0$. But then the coefficient of $\alpha_3$ is also zero, a contradiction. Similar arguments show that no contrasts involving $\alpha_3$ or $\beta_3$ are estimable.

3. Suppose that data $(x_i, Y_{ij})$, $i = 1, \ldots, k,\ j = 1, \ldots, n_i$, are available and the goal of the analysis is to fit a regression model

$$Y_{ij} = m(x_i) + \varepsilon_{ij}$$

to the data. The error terms are assumed to be independent with a common $N(0, \sigma^2)$ distribution, and at least one of the $n_i$ is greater than one.

(a) What statistic would you use to test $H_0$: $m(x) = \beta_0 + \beta_1 x$ against the general alternative? This is the test of lack of fit. Give the formula for the test statistic and its distribution under $H_0$. You do not need to provide a derivation.

(b) If $n_i = 1$ for each $i$, how would you decide whether a straight line model fits the data?

(c) Suppose that the data are given in the following table:

| $x$ | 60 | 70 | 70 | 80 | 80 | 90 | 90 | 90 | 100 | 100 | 110 | 110 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 51 | 82 | 78 | 91 | 96 | 98 | 89 | 99 | 82 | 83 | 54 | 52 |

A linear regression equation was fitted to the data, and the following partial ANOVA table was computed.

| Source | Sum of Squares | d.f. | Mean Square |
|---|---|---|---|
| Linear regression | 69.31 | ? | ? |
| Lack of fit | ? | ? | ? |
| Error | 83.67 | ? | ? |
| Corrected total | 3462.91 | 11 | |

The error sum of squares in the table is $\sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2$. Complete the missing entries in the ANOVA table. Does the linear model seem to fit the data? How might you improve on the model?

**Solution.** (a) Write $N = \sum_i n_i/N$ , $\bar{Y} = \sum_i \sum_j Y_{ij}/N$ and $\bar{x} = \sum_i n_i x_i/N$. Under $H_0$ we have a linear regression with least squares estimates

$$\hat{\beta}_1 = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (x_i - \bar{x}) Y_{ij}}{\sum_{i=1}^k n_i (x_i - \bar{x})^2}; \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}.$$

Under the alternative $\widehat{m}(x_i) = \bar{Y}_{i\cdot}$. Therefore the lack of fit test rejects for large values of

$$\mathcal{F} = \frac{(\bar{Y}_{i\cdot} - \bar{Y} - \hat{\beta}_1 (x_i - \bar{x}))^2/(k-2)}{\sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2/(N-k)},$$

which has the $F_{k-2,N-k}$ distribution if $H_0$ is true.

(b) If no replicated observations are present, one can fit a more complex linear model and test whether the linear model fits as well as the complex model. For instance one might fit a low degree polynomial to the data and test whether the higher degree terms have coefficient zero. One can also examine residual plots and look for substantial departures from a "white noise" pattern.

(c) The completed ANOVA table is the following.

| Source | Sum of Squares | d.f. | Mean Square |
|---|---|---|---|
| Linear regression | 69.31 | **1** | **69.31** |
| Lack of fit | **3309.93** | **4** | **827.48** |
| Error | 83.67 | **6** | **13.95** |
| Corrected total | 3462.91 | 11 | |

The $F$ statistic for lack of fit is $827.48/13.95 = 59.34$, so we conclude that the linear model does not fit the data. Judging from the pattern in the data (large $Y$ values in the middle and small values at the extremes of $x$), a quadratic term would greatly improve the fit.

4. Consider the quadratic regression model

$$Y_{ij} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + e_{ij}$$

where $n$ observations of $Y_{ij}$ are made for each of the following combinations of $(x_{i1}, x_{i2})$: $(1,1)$, $(1,-1)$, $(-1,1)$, $(-1,-1)$ and $(0,0)$.

(a) Assume $n = 1$ and $\beta_{11} = \beta_{22} = 0$. Write the model in matrix form. Which parameters, if any, are estimable?

(b) Under the assumptions of (a), compute a set of least squares estimates. How would your answers change if $n > 1$?

(c) If $n > 1$ and we make no assumptions about $\beta_{11}$ and $\beta_{22}$, how would you estimate $\sigma^2 = \mathrm{Var}\,(e_{ij})$?

(d) Can you test $H_0$: $\beta_{11} + \beta_{22} = 0$? You do not need to construct the test statistic, but describe how it could be constructed and state its distribution under $H_0$.

**Solution.** (a) In matrix form with $n = 1$ we have

$$\mathbf{Y} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_{12} \end{bmatrix} + \mathbf{e}$$

Note that $\mathbf{X}$ has full rank since the columns are orthogonal. Therefore all parameters are estimable.

(b) It is easy to see that

$$\mathbf{X}^T\mathbf{X} = \mathrm{diag}(5,4,4,4); \qquad \mathbf{X}^T\mathbf{Y} = \begin{bmatrix} Y_1 + Y_2 + Y_3 + Y_4 + Y_5 \\ Y_1 + Y_2 - Y_3 - Y_4 \\ Y_1 - Y_2 + Y_3 - Y_4 \\ Y_1 - Y_2 - Y_3 + Y_4 \end{bmatrix}.$$

The least squares solution is

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} (Y_1 + Y_2 + Y_3 + Y_4 + Y_5)/5 \\ (Y_1 + Y_2 - Y_3 - Y_4)/4 \\ (Y_1 - Y_2 + Y_3 - Y_4)/4 \\ (Y_1 - Y_2 - Y_3 + Y_4)/4 \end{bmatrix}.$$

With $n$ replicated observations at each $(x_1, x_2)$ combination, the rows of the above $\mathbf{X}$ are repeated $n$ times and $\mathbf{X}^T\mathbf{Y}$ is multiplied by $n$. To get the estimators, replace $Y_i$ by $\bar{Y}_{i\cdot}$ in the formula above.

(c) With $\beta_{11}$ and $\beta_{22}$ in the model, we have $p = 6$ and $r = 5$ because $x_1^2 \equiv x_2^2$. One can still solve the least squares equations and compute a sum of squared residuals. Then $s^2 = SSE/(5n - 5)$ is an unbiased estimator. [In fact, $SSE = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2$.]

(d) The hypothesis $H_0$: $\beta_{11} + \beta_{22} = 0$ is testable because $\beta_{11} + \beta_{22}$ is the coefficient of $x_1^2 \equiv x_2^2$ in the model. We can solve the least squares equations (under the side condition $\beta_{22} = 0$) and use the statistic $t = \hat{\beta}_{11}/\text{se}(\hat{\beta}_{11})$ to test $H_0$. We could also look at the reduction in the sum of squares due to fitting the full model vs. the model of (a).