# Applied Statistics Preliminary Examination
# Theory of Linear Models
# May 2018

**Instructions:**

- Do all 3 Problems. Neither calculators nor electronic devices of any kind are allowed. Show all your work, clearly stating any theorem or fact that you use. Each of the 15 parts carries an equal weight of 10 points.
- Abbreviations/Acronyms.
  - IID (independent and identically distributed).
  - LSE (least squares estimator); BLUE (best linear unbiased estimator). Sometimes the LSE may be designated OLS (ordinary least squares) estimator, in order to differentiate it from the GLS (generalized least squares) estimator.
- Notation.
  - $\boldsymbol{x}^T$ or $\boldsymbol{A}^T$: indicates transpose of vector $\boldsymbol{x}$ or matrix $\boldsymbol{A}$.
  - $\mathrm{tr}(\boldsymbol{A})$ and $|\boldsymbol{A}|$: denotes the trace and determinant, respectively, of matrix $\boldsymbol{A}$.
  - $\boldsymbol{I}_n$: the $n \times n$ identity matrix.
  - $\boldsymbol{j}_n = (1, \ldots, 1)^T$ is an $n$-vector of ones, and $\boldsymbol{J}_{m,n}$ is an $m \times n$ matrix of ones.
  - $\mathbb{E}(X)$ and $\mathrm{var}(X)$: expectation and variance of random variable $X$.
  - $\boldsymbol{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: the $m$-dimensional random vector $\boldsymbol{x}$ has a normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
  - $X \sim t(n, \lambda)$: a $t$ distribution with $n$ degrees of freedom and noncentrality parameter $\lambda$. If $\lambda = 0$ we write simply: $X \sim t(n)$.
  - $X \sim F(n_1, n_2, \lambda)$: an $F$ distribution with $n_1$ and $n_2$ numerator and denominator degrees of freedom respectively, and noncentrality parameter $\lambda$. If $\lambda = 0$ we write simply: $X \sim F(n_1, n_2)$.
- Possibly useful results.
  - If $X \sim F(p, q)$, then recall that $\mathbb{E}(X) = q/(q-2)$ for $q > 2$, and note the following result:

$$\frac{\left(\frac{p}{q}\right) X}{1 + \left(\frac{p}{q}\right) X} \sim \mathrm{Beta}\left(\frac{p}{2}, \frac{q}{2}\right).$$

    Also recall that if $Y \sim \mathrm{Beta}(a, b)$, then $\mathbb{E}(Y) = a/(a+b)$.

1. Consider the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim (\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a positive definite covariance matrix, and the $n \times k$ design matrix $\boldsymbol{X}$ is of full (column) rank. Let $\hat{\boldsymbol{\beta}}_{OLS}$ and $\hat{\boldsymbol{\beta}}_{GLS}$ denote the OLS and GLS estimators of $\boldsymbol{\beta}$, respectively, and recall that the latter is the BLUE in this case.

   (a) Compute $\text{var}(\hat{\boldsymbol{\beta}}_{OLS})$ and $\text{var}(\hat{\boldsymbol{\beta}}_{GLS})$ under the model.

   (b) Show that $\text{cov}(\hat{\boldsymbol{\beta}}_{OLS}, \hat{\boldsymbol{\beta}}_{GLS}) = \text{var}(\hat{\boldsymbol{\beta}}_{GLS})$.

   (c) Compute $\text{cov}(\hat{\boldsymbol{\beta}}_{GLS}, \hat{\boldsymbol{\beta}}_{GLS} - \hat{\boldsymbol{\beta}}_{OLS})$.

   (d) Using the above results, show, without appealing to the Gauss-Markov Theorem, that $\text{var}(\hat{\boldsymbol{\beta}}_{OLS}) - \text{var}(\hat{\boldsymbol{\beta}}_{GLS})$ is a positive semidefinite matrix.

   (e) If the columns of $\boldsymbol{X}$ consist of a set of $k$ orthonormal eigenvectors of $\boldsymbol{\Sigma}$, compute the resulting estimators $\hat{\boldsymbol{\beta}}_{OLS}$ and $\hat{\boldsymbol{\beta}}_{GLS}$. Are they the same?

2. Consider the usual linear model, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where the $n \times (k+1)$ matrix $\boldsymbol{X}$ is of full rank with first column equal to $\boldsymbol{j}_n$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T$, and $\boldsymbol{\epsilon} \sim \text{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Letting $\hat{\boldsymbol{\beta}}$ denote the usual LSE of $\boldsymbol{\beta}$, recall the definition of the *coefficient of determination*:

$$R^2 = \frac{(\boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{y} - n\overline{y}^2}{\boldsymbol{y}^T\boldsymbol{y} - n\overline{y}^2}, \qquad \text{where} \quad \overline{y} = \frac{1}{n}\boldsymbol{j}_n^T\boldsymbol{y}.$$

   This question will consider the properties of $R^2$ under the test of *overall regression*, corresponding to the null hypothesis, $H_0 : \beta_1 = \cdots = \beta_k = 0$.

   (a) With $N = n - k - 1$, show that the $F$-statistic for testing $H_0$ can be expressed as: $F = \frac{R^2/k}{(1-R^2)/N}$.

   (b) Determine the distribution of the $F$-statistic in (a) under $H_0$.

   (c) Show that under $H_0$, $\mathbb{E}(1/R^2) = (n-3)/(k-2)$.

   (d) Show that under $H_0$, $R^2 \sim \text{Beta}(a, b)$, and determine the values of $a$ and $b$.

   (e) Hence, or otherwise, show that under $H_0$, $\mathbb{E}(R^2) = k/(n-1)$.

3. Consider the vector of observations $\boldsymbol{y} = (y_1, \ldots, y_4)^T$ from the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, given in expanded form as:

$$\boldsymbol{y} = \begin{pmatrix} 1 & 1 & 1 & -1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n).$$

   (a) Determine the rank of $\boldsymbol{X}$, and hence show that $\boldsymbol{X}$ is *rank-defficient*. (From now on assume that a generalized inverse, $\boldsymbol{G}$, of $\boldsymbol{X}^T\boldsymbol{X}$ is available.)

   (b) Determine which of the following are *estimable* functions, and for the estimable ones, find their BLUEs and their distributions:

$$\beta_2, \qquad \beta_4, \qquad 2\beta_1 + \beta_3, \qquad \beta_2 + \beta_4, \qquad \beta_2 - \beta_4.$$

   (c) Show that the null $H_0 : \{4\beta_1 + 2\beta_3 = 0,\ \beta_2 = \beta_4\}$ is a *testable* hypothesis, and construct a test statistic for it.

   (d) Find the distribution the test statistic in (c), both under $H_0$ as well as under the alternative $H_1 : \{4\beta_1 + 2\beta_3 = 2,\ \beta_2 = \beta_4 + 2,\ \sigma^2 = 1/2\}$.

   (e) Specifically determine a *reparametrization* of the model to full rank, where $\boldsymbol{\gamma}$ is the new parameter vector in the full rank linear model $\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$. That is, indicate how to compute the new (full rank) design matrix $\boldsymbol{Z}$, and express $\boldsymbol{\gamma}$ as a function of $\boldsymbol{\beta}$.

# Design of Experiment: Prelim Problems
## May 2018
## Please Do All Problems. Each of the 14 parts carries an equal weight of 10 points.

1. An experiment described by Johnson and Leone (1977, p. 758) was performed by a company to investigate the effects of various factors on the yield strength of a particular titanium alloy. The factors investigated were:

   A vendors (4 fixed levels representing suppliers of raw material).

   C bar size (2 fixed levels representing standard sizes of bars of raw material).

   B batch (3 randomly selected levels nested within each combination of levels of A and C).

   D product type (2 fixed levels representing different types of finished product forgedown and finished-forge blades).

   Three observations were taken on each of ABCD treatment combination. A reasonable model was thought to have all the main effects and AC, AD, and BD interactions.

   a) Please write down the model, and explain all the notations and assumptions in the model.

   b) Write down the degrees of freedom and expected mean squares column of the analysis of variance table.

   c) Show that the expected mean squares for A is the same as found in b).

   d) Give an unbiased estimation for $\sigma^2_{B(AC)}$, and give a formula for a 95% confidence interval for $\sigma^2_{B(AC)}$.

   e) How would you test the following hypothesis?

   $H_0$ : {no differences in yield strength of the titanium alloy can be attributed to the four vendors}

   $H_A$ : {$H_0$ is false}

2. Suppose that factors C and D are to be investigated further in a followup experiment. Suppose that two new factors P and Q (heat setting during processing and cooling method) are also to be investigated at two levels each. A followup experiment is required with the four factors C, D, P, and Q at two levels each (a $2^4$ experiment). Only sixteen observations will be taken, four for each vendor. It is known that the interactions CP, CQ, PQ, CPQ, and CDPQ are likely to be negligible. Also, there was information gained from the previous parts to Question 1 to suggest that all interactions of treatment factors with vendor can be assumed negligible.

   a) Divide the 16 treatment combinations into four blocks of size four (one block for each vendor).

   b) Write down a suitable model and the degrees of freedom column for the analysis of variance table for your design in part a).

3. Engineers performed an experiment to investigate warping of copper plates. The two factors studied were the temperature (°C), that the plates were exposed to, and the copper content (%) of the plates. The response variable was a measure of the amount of warping. Four temperatures (50, 75, 100, 125) and four copper contents(40%, 60%, 80%, 100%) were investigated. A completely randomized design was used and there were two measurements on each of the 16 possible temperature-copper combinations.

   a) Write down a two-factor analysis of variance model for analyzing these data. Be sure to indicate what any symbols you use mean and any assumptions you make.

   b) The data were analyzed using software. Table 1 and 2 are partial outputs from the analysis of variance and some descriptive statistics. Please fill in the missing terms.

Table 1: ANOVA

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Copper | _____ | _____ | _____ | _____ | _____ |
| Temp | _____ | 156.094 | _____ | 7.67 | 0.002 |
| Copper∗Temp | _____ | 113.781 | 12.642 | 1.86 | 0.133 |
| Error | _____ | _____ | _____ | | |
| Total | _____ | 1076.719 | | | |

Table 2: Descriptive Statistics

| Variable | Copper | N | Mean |
|---|---|---|---|
| Warp | 40 | 8 | 15.5 |
| | 60 | 8 | 18.88 |
| | 80 | 8 | 21.00 |
| | 100 | 8 | 28.250 |
| Overall Mean | | | 20.9075 |

   c) Figure 1 is an interaction plot created from software. The profiles in the plot are not parallel, suggesting an interaction may be present. Yet the ANOVA table indicates that there is not a statistically significant interaction at the 0.05 level. Does this suggest that an error is present? Explain.

   d) Figure 2 is a plot of the residual versus fitted values. Does this plot indicate that there may be problems with some of our ANOVA assumptions? In particular, one of the engineers noticed the plot does not look like a collection of randomly scattered points, so there might be problems with the assumption of normality. Comment.

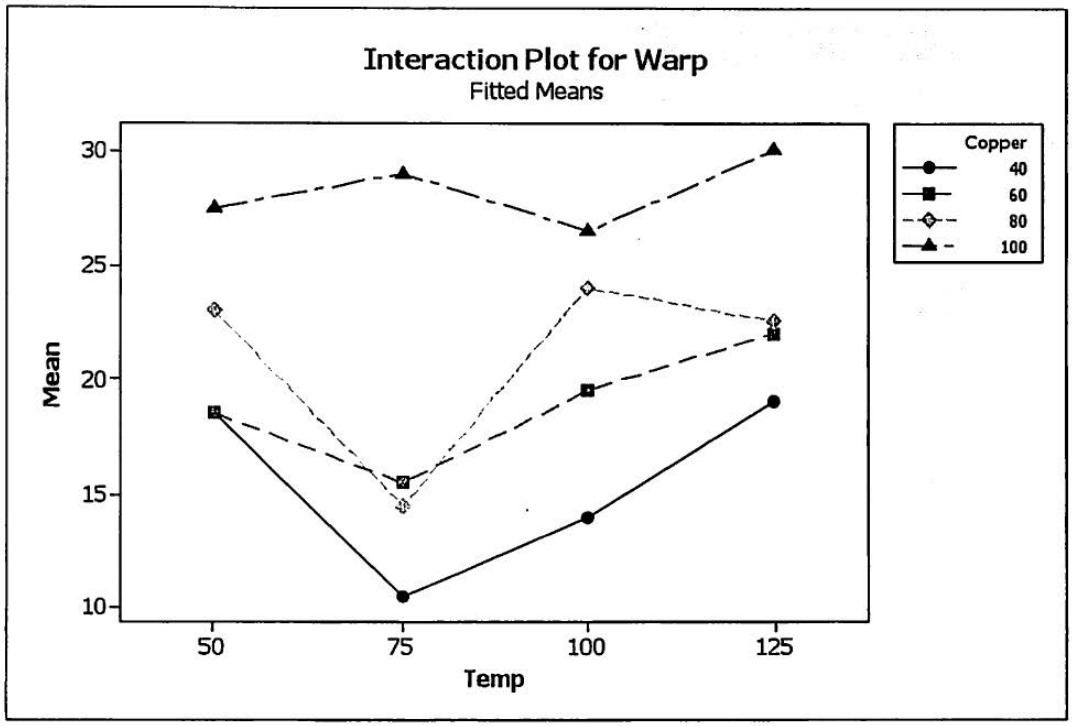


Figure 1: Interaction Plot

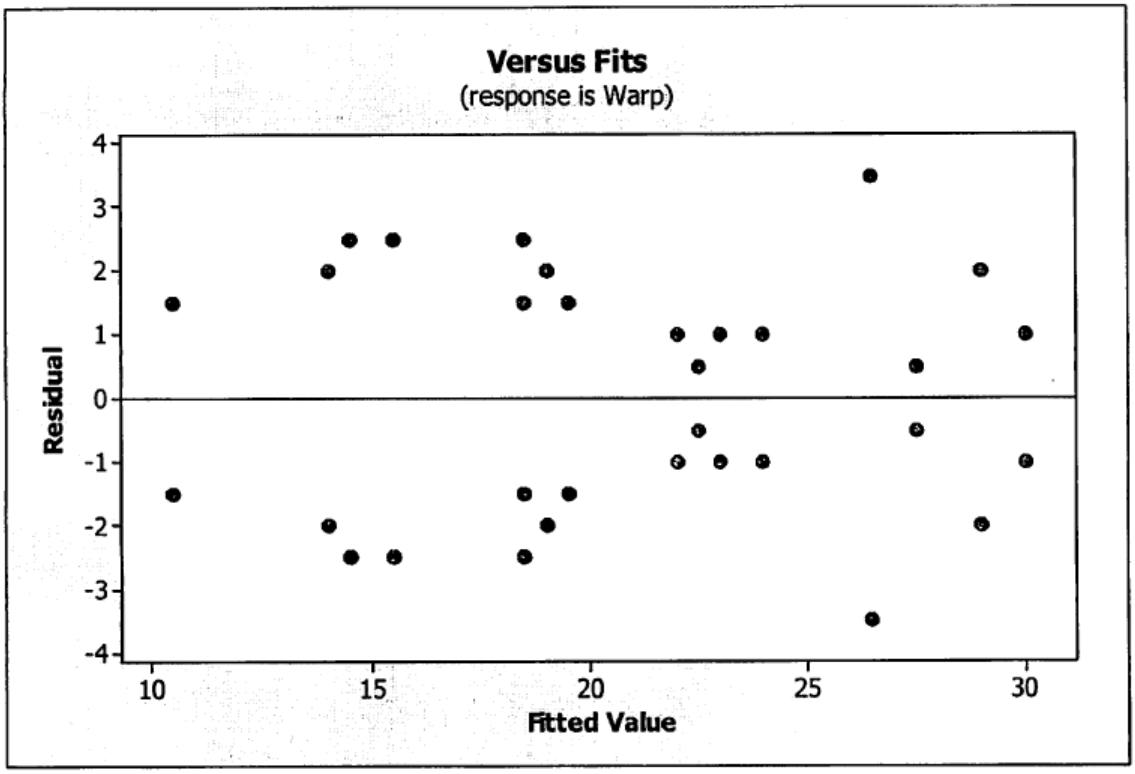


Figure 2: Residual vs. fitted values

e) If the object is to minimize warping across all temperatures, what level of copper content would you recommend? Construct overall 90% confidence intervals to justify your answer.

f) If one wants the width of overall 90% confidence interval for pairwise comparing warp values across different copper groups to be at most 2 units (msd=1), then please write down an inequation to identify proper sample size, which is an expression of the number of replication $r$ for each treatment combination. You don't need to do the calculation.

g) If in practice the temperature is very hard to change, please give suggestions about what type of design the experimenter might want to use. Briefly describe the revised design and analysis model. Which main effect do you expect has more accurate estimation?