Ph.D. PRELIMINARY EXAMINATION

March 20, 2003

PART II: Methods

Consider the following model for a one-way classification with one concomitant variable:

$$y_{ij} = \mu_i + \gamma z_{ij} + \epsilon_{ij}, \quad i = 1, \ldots, t, \quad j = 1, \ldots, s,$$

where $t \geq 3$, $s \geq 2$, the μ_i's and γ are unknown fixed parameters, the y_{ij}'s are observable variables, and the z_{ij}'s are known constants. Here, the ϵ_{ij}'s are independently distributed as $N(0, \sigma^2)$, where σ^2 is an unknown positive parameter.

Let $\underline{\beta} = (\mu_1, \ldots, \mu_t, \gamma)'$. Let

$$\overline{y}_{i\cdot} = \sum_{j=1}^{s} y_{ij}/s, \quad \overline{z}_{i\cdot} = \sum_{j=1}^{s} z_{ij}/s,$$

and

$$m_{yy} = \sum_{i=1}^{t}\sum_{j=1}^{s}(y_{ij} - \overline{y}_{i\cdot})^2, \quad m_{zz} = \sum_{i=1}^{t}\sum_{j=1}^{s}(z_{ij} - \overline{z}_{i\cdot})^2, \quad m_{yz} = \sum_{i=1}^{t}\sum_{j=1}^{s}(y_{ij} - \overline{y}_{i\cdot})(z_{ij} - \overline{z}_{i\cdot}).$$

Note: You may use matrices in your derivations, but no matrix should appear in your final answers to the following questions except for part (a).

(a) Formulate the model as a Gauss-Markov model:

$$\underline{y} = X\underline{\beta} + \underline{\epsilon}, \quad E(\underline{\epsilon}) = \underline{0}, \quad \mathrm{Var}(\underline{\epsilon}) = \sigma^2 I.$$

Carefully write out $\underline{y}$, X, and $\underline{\epsilon}$.

(b) Give necessary and sufficient conditions on the z_{ij}'s for X to be of full rank.

For the following questions, assume that X is of full rank.

(c) Show that the best linear unbiased estimator of $\underline{\beta}$ is given by $\widehat{\underline{\beta}} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_t, \widehat{\gamma})'$, where

$$\widehat{\gamma} = m_{yz}/m_{zz}, \quad \widehat{\mu}_i = \overline{y}_{i\cdot} - \widehat{\gamma}\overline{z}_{i\cdot}.$$

(d) Give an unbiased estimator, $\widehat{\sigma}^2$, of σ^2.

(e) Show that $\widehat{\gamma}$ (given in part (c)) and the $\overline{y}_{i\cdot}$'s are statistically independent.

(f) Show that the distribution of $\widehat{\gamma}$ is given by $N(\gamma, \sigma^2/m_{zz})$.

(g) Derive a $(1-\alpha)$ confidence interval for γ.

(h) Derive a size-α test for testing $H_0 : \mu_1 = \cdots = \mu_t$ versus H_a : not H_0.

(i) Use two different methods and give confidence intervals for all the differences $\mu_i - \mu_j$ $(i, j = 1, \ldots, t, i \neq j)$ such that the probability of simultaneous coverage is at least $1 - \alpha$.

Ph.D. Prelim Exam
Spring 2003

Solutions
Linear Models

(a) Let $\underline{1}_s$ be an $s \times 1$ vector of 1's, and $\underline{z}_i = (z_{i1}, \cdots, z_{is})'$, for $i = 1, \cdots, t$. Then

$$\underline{y} = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1s} \\ \vdots \\ y_{t1} \\ \vdots \\ y_{ts} \end{bmatrix}, \quad X = \begin{bmatrix} \underline{1}_s & \underline{0} & \cdots & \underline{0} & \underline{z}_1 \\ \underline{0} & \underline{1}_s & \cdots & \underline{0} & \underline{z}_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \underline{0} & \underline{0} & \cdots & \underline{1}_s & \underline{z}_t \end{bmatrix}, \text{ and } \underline{\varepsilon} = \begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1s} \\ \vdots \\ \varepsilon_{t1} \\ \vdots \\ \varepsilon_{ts} \end{bmatrix}.$$

(b) Note that X is of full rank $(t+1)$ if and only if for at least one $i \in \{1, \cdots, t\}$, there exist $j, k \in \{1, \cdots, s\}$ such that $z_{ij} \neq z_{ik}$.

(c) Note that $\underline{\beta}$ is an estimable vector and its best linear unbiased estimator is the least squares estimator of $\underline{\beta}$ (since we have a Gauss-Markov model).

Further note that

$$\min_{\underline{\beta}} \sum_{i,j} (y_{ij} - \mu_i - \gamma z_{ij})^2$$

$$= \min_{\gamma} \min_{\mu_1, \cdots, \mu_t} \sum_i \sum_j (y_{ij} - \mu_i - \gamma z_{ij})^2.$$

Consider for a fixed γ; $\sum_i \sum_j (y_{ij} - \mu_i - \gamma z_{ij})^2$

is minimized when $\mu_i = \bar{y}_{i\cdot} - \gamma \bar{z}_{i\cdot}$. Thus

$$\min_{\beta} \sum_i \sum_j (y_{ij} - \mu_i - \gamma z_{ij})^2$$

$$= \min_{\gamma} \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot} - \gamma (z_{ij} - \bar{z}_{i\cdot}))^2.$$

Solving the above minimization problem gives $\hat{\gamma} = \frac{m_{yz}}{m_{zz}}$ and thus

$$\hat{\mu}_i = \bar{y}_{i\cdot} - \hat{\gamma} \cdot \bar{z}_{i\cdot}.$$

(d) An unbiased estimator of σ^2 is given by

$$\frac{\sum_i \sum_j (y_{ij} - \hat{\mu}_i - \hat{\gamma} z_{ij})^2}{t \cdot s - t - 1} = \frac{1}{ts - t - 1} \cdot \left(m_{yy} - \frac{m_{yz}^2}{m_{zz}}\right).$$

(e) Note that the joint distribution of $\hat{\gamma}$ and the $\bar{y}_{i\cdot}$'s is multivariate normal. Thus it suffices to show that $\hat{\gamma}$ and $\bar{y}_{i\cdot}$ are uncorrelated. Since y_{kj}'s $(k \neq i)$ and $\bar{y}_{i\cdot}$ are clearly uncorrelated, it suffices to show that for each j, $y_{ij} - \bar{y}_{i\cdot}$ and $\bar{y}_{i\cdot}$ are uncorrelated, which is indeed true by observing

that $\operatorname{cov}(y_{ij}-\bar{y}_{i\cdot}, \bar{y}_{i\cdot}) = \operatorname{cov}(y_{ij}, \bar{y}_{i\cdot}) - \operatorname{Var}(\bar{y}_{i\cdot})$
$= \sigma^2/s - \sigma^2/s = 0.$

(f) Note that $\hat{r}$ is an unbiased estimator of r and is normally distributed. Thus

$$\hat{r} \sim N(r, \operatorname{Var}(\hat{r})).$$

To compute $\operatorname{Var}(\hat{r})$, consider for any constants $a_1, \cdots, a_s$ and independent $N(\mu, \sigma^2)$ random variables $u_1, \cdots, u_s$, and $\bar{u} = (\sum_{j=1}^{s} u_j)/s$,

$$\operatorname{Var}\left(\sum_{j=1}^{s} a_j(u_j - \bar{u})\right) = \operatorname{Var}\left(\sum_{j=1}^{s} a_j u_j - \left(\sum_{j=1}^{s} a_j\right)\cdot \bar{u}\right)$$

$$= \operatorname{Var}\left(\sum_{j=1}^{s} a_j u_j\right) - 2\operatorname{Cov}\left(\sum_{j=1}^{s} a_j u_j, \left(\sum_{j=1}^{s} a_j\right)\bar{u}\right) + \operatorname{Var}\left(\sum_{j=1}^{s} a_j \bar{u}\right)$$

$$= \left(\sum_{j=1}^{s} a_j^2\right)\sigma^2 - 2\left(\sum_{j=1}^{s} a_j\right)^2 \cdot \frac{\sigma^2}{s} + \left(\sum_{j=1}^{s} a_j\right)^2 \cdot \frac{\sigma^2}{s}$$

$$= \sum_{j=1}^{s} (a_j - \bar{a})^2 \cdot \sigma^2, \quad \text{where } \bar{a} = \sum_{j=1}^{s} a_j/s.$$

Thus $\operatorname{Var}(\hat{r}) = \frac{1}{m_{zz}^2} \sum_{i=1}^{t} \operatorname{Var}\left(\sum_{j=1}^{s} (y_{ij} - \bar{y}_{i\cdot})(z_{ij} - \bar{z}_{i\cdot})\right)$

$$= \frac{1}{m_{zz}^2} \underbrace{\sum_{i=1}^{t} \sum_{j=1}^{s} (z_{ij} - \bar{z}_{i\cdot})^2}_{\| \; m_{zz}} \cdot \sigma^2 = \frac{\sigma^2}{m_{zz}}.$$

(g) Let $\hat{\sigma}^2$ be defined in part (d). Then

$$\frac{\hat{\gamma}-\gamma}{\hat{\sigma}/\sqrt{m_{zz}}} \sim t_{ts-t-1}.$$

Thus a $(1-\alpha)$ confidence interval for γ is given by $\hat{\gamma} \pm t_{\alpha/2:(ts-t-1)} \cdot \frac{\hat{\sigma}}{\sqrt{m_{zz}}}$, where $\hat{\gamma} = m_{yz}/m_{zz}$ and $t_{\alpha/2:(ts-t-1)}$ is the upper $(\alpha/2)$ point of a t distribution with $(ts-t-1)$ degrees of freedom.

(h) Let $\bar{y}_{..} = \sum_{i=1}^{t}\sum_{j=1}^{s} y_{ij}/(ts)$, $\bar{z}_{..} = \sum_{i=1}^{t}\sum_{j=1}^{s} z_{ij}/(ts)$,

$$\tilde{m}_{yy} = \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2, \quad \tilde{m}_{yz} = \sum_i \sum_j (y_{ij} - \bar{y}_{..})\cdot(z_{ij} - \bar{z}_{..}),$$

and $\tilde{m}_{zz} = \sum_i \sum_j (z_{ij} - \bar{z}_{..})^2$. Using the method in part (c), we obtain the least squares estimator of β under the given model with $\mu_1 = \cdots = \mu_t$. The estimator is $\tilde{\beta} = (\tilde{\mu}_1, \cdots, \tilde{\mu}_1, \tilde{\gamma})'$, where $\tilde{\mu}_1 = \bar{y}_{..} - \tilde{\gamma}\cdot\bar{z}_{..}$ and $\tilde{\gamma} = \tilde{m}_{yz}/\tilde{m}_{zz}$. Thus, the residual sum of squares under the reduced model (with $\mu_1 = \cdots = \mu_t$) is

$$\tilde{m}_{yy} - \frac{\tilde{m}_{yz}^2}{\tilde{m}_{zz}}.$$

It follows from the general ANOVA table and Cochran's theorem that

$$F = \frac{\left(\tilde{m}_{yy} - \frac{\tilde{m}_{yz}^2}{\tilde{m}_{zz}} - m_{yy} + \frac{m_{yz}^2}{m_{zz}}\right)/(t-1)}{\left(m_{yy} - \frac{m_{yz}^2}{m_{zz}}\right)/(ts-t-1)}$$

has a central $F_{t-1,\,ts-t-1}$ distribution under H_0. Thus a size-α test of $H_0: \mu_1 = \cdots = \mu_t$ versus H_a: not H_0 is obtained by rejecting H_0 if and only if $F \geq F_{\alpha:(t-1),(ts-t-1)}$.

(i) The simultaneous confidence intervals for all the differences $\mu_i - \mu_j$ can be obtained by the following two methods:

(1) Scheffé's method:

$$(\hat{\mu}_i - \hat{\mu}_j) \pm \sqrt{(t-1)\,F_{\alpha:(t-1),(ts-t-1)} \cdot \hat{\sigma}^2 \cdot \left[\frac{2}{s} + \frac{(\bar{z}_{i\cdot} - \bar{z}_{j\cdot})^2}{m_{zz}}\right]}$$

(using the results from parts (c) — (f)).

(2) The Bonferroni method:

$$(\hat{\mu}_i - \hat{\mu}_j) \pm t_{\alpha/(t(t-1)):(ts-t-1)}\ \hat{\sigma} \cdot \sqrt{\frac{2}{s} + \frac{(\bar{z}_{i\cdot} - \bar{z}_{j\cdot})^2}{m_{zz}}}.$$

Note that Tukey's method is not applicable here.

Photosynthesis is the process of converting CO_2 from the air to sugars that are stored or used in a plant. It occurs in plant leaves in the presence of light. At the same time, the process of respiration converts sugars in the leaf to CO_2. The net photosynthetic rate (NPR) is the difference in the rates of the two processes (photosynthesis - respiration). NPR is positive when a plant is in light and negative when the plant is in the dark. A simple model for the relationship between light intensity and NPR is the linear saturation model illustrated below. The dark respiration rate is the NPR when the light intensity = 0. NPR increases linearly with light intensity until a light saturation point, beyond which NPR is constant. Physiological ecologists are interested in three values:

- α The dark respiration rate (i.e. NPR when light intensity = 0)
- β The slope of the linear portion of the photosynthesis curve
- γ The light saturation point (i.e. the light intensity at which NPR no longer increases).

These three parameters, and values derived from them, are useful to compare photosynthetic performance among plants or among species. Figure 1 shows the typical relationship between NPR and light intensity and illustrates important quantities.

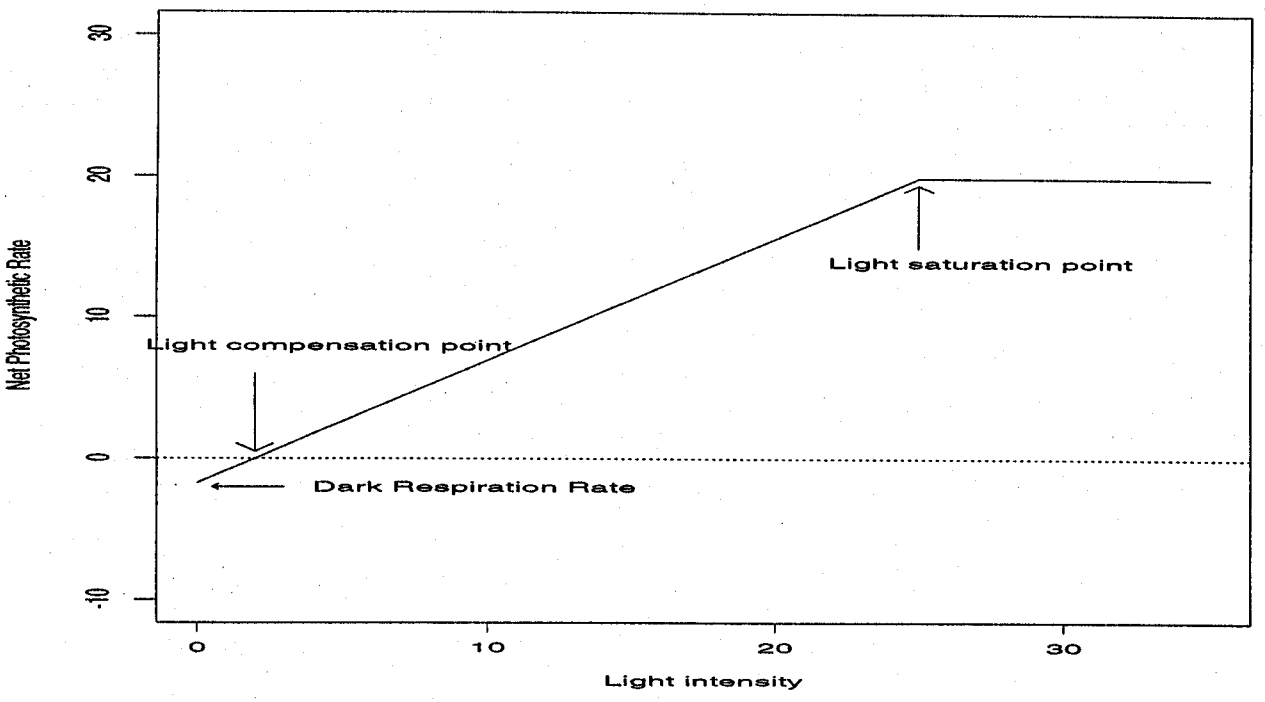


Figure 1: Linear saturation model for net photosynthesis

This problem looks at various questions and analyses for these curves. There are five parts; they are generally independent of each other. If you get stuck on one part, skip to another.

Part A (A-1) Conditional on the light intensity and parameters, it is reasonable to assume that observed NPR is normally distributed. Write out a model for **mean** NPR that describes patterns like in figure 1 in terms of the three parameters defined above, dark respiration, slope, and light saturation point. Be sure to define your symbols if you use any symbols not defined above.

(A-2) A group of investigators has collected data from one plant, measured at 9 different light intensities.

Assume: γ, the light intensity at the saturation point is a fixed, known constant, and σ^2, the variance of observations around the line is constant

Identify a reasonable and simple method to estimate the unknown parameters (α, β, and σ^2). You do not have to provide estimators, but you need to describe your method in sufficient detail that someone else could repeat your analysis.

(A-3) The investigators also have data for a plant of a different species. The light saturation point, γ, for this species is unknown and must be estimated. Can your method from part A-2 be used when the light saturation point is unknown? If not, describe a reasonable method to estimate all the parameters in your model, including the light saturation point. Again, you don't need to provide estimators, but you need to provide sufficient detail so that someone else could repeat your analysis.

Part B Besides the three quantities listed above (dark respiration, slope, and light saturation point), physiological ecologists are very interested in the light compensation point. This is the light intensity at which photosynthesis = respiration, i.e. NPR = 0. This point is indicated on figure 1.

The model in part A-1 was fit to data from the single plant used in part A-2. The estimates and their variance-covariance matrix are:

Coefficient	Estimate	Variance-Covariance Matrix	
Intercept, α	-0.9732	0.5671	-0.02748
Slope, β	0.8523	-0.02748	0.001940

(B-1) Estimate the light intensity at the compensation point.

(B-2) Estimate the standard error of the light intensity at the compensation point.

Part C (C-1) This part is a detailed look at data collected on a single plant. NPR was measured five times at each of nine light intensities. The light intensities were used in a random order, with sufficient time between measurements to eliminate any correlation between observations. Figure 2a is a plot of the observations; figure 2b is a plot of residuals and predicted values. Do you have any concerns about your method in part A-2?

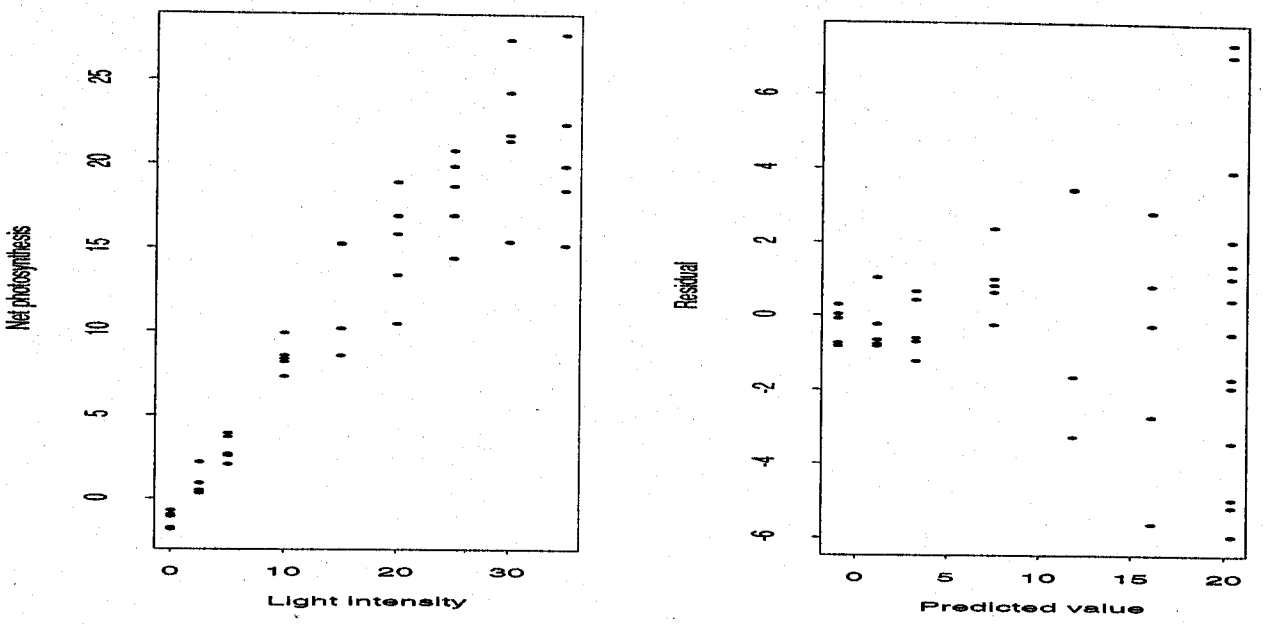


Figure 2: Plots of a) observations and b) residuals vs. predicted values, for data from one plant

(C-2) The investigators want to model the relationship between the variance in NPR and the light intensity. Remember, NPR is the difference between the photosynthetic rate and the respiration rate. The respiration rate is not affected by the light intensity. Respiration rate is a random variable with a constant mean and constant variance. The rate of photosynthesis (**not NPR**, just photosynthesis) is a random variable with constant coefficient of variation. The two random variables can be assumed to be independent. Construct a model for the variance of NPR as a function of light intensity and unknown parameters. Be sure to define your parameters.

(C-3) Describe a method to evaluate whether the data are consistent with variance model from part C-2.

(C-4) Assume that the variance of observations follows the variance model from part C-2, although the parameters of that model are unknown. Describe a reasonable method to estimate parameters in the mean model (from part A-1) and the parameters in the variance model (from part C-2). Again, assume that the light intensity at the saturation point is a known constant. You do not need to provide estimators, but you do need to describe your method in sufficient detail for someone else to repeat the analysis.

Part D (D-1) The data used in part C were collected from a single plant. The investigators have data from a second plant of the same species. They would like to know if the two plants have the same light response curve. They don't care whether the dark respiration (α) differs, the slope (β) differs, or whether both differ. One way to test this is to compare models. Residual Sums-of-squares for various models are:

Intercept, α	Slope, β	Residual SS
Same	Same	75.76
Same	Different	72.75
Different	Same	75.51
Different	Different	69.93

There are a total of 90 observations (2 plants x 9 light levels x 5 reps). All models were fit using the appropriate variance structure; all models assume that the light saturation points are the same.

If possible, tell the investigators whether the second plant has a different light response curve. Provide your test statistic, the p-value, and a one sentence answer in words. If this test is not possible with the quantities given, indicate what statistics or other information is needed.

(D-2) The previous analyses assume that there is a light saturation point at a known constant. The investigators tested whether there is a light saturation point by comparing the fit of the model from part A to the fit of a linear regression without any saturation point ($\mathrm{E}\ Y = \alpha + \beta X$). Using the appropriate variance models, the residual sums-of-squares are:

Model	Residual SS
Saturation Point	28.38
Linear Regression	38.11

If possible, tell the investigators whether this plant has a light saturation point. Provide your test statistic, the p-value, and a one sentence answer in words. If this test is not possible with the quantities given, indicate what statistics or other information is needed.

Part E The ultimate goal of the research is to compare NPR curves for two species. The light saturation points are not known for either species. The investigators have measured 10 plants from each species, a total of 20 plants. Each plant was measured twice at 9 different light levels, giving a total of 360 observations (2 species x 10 plants / species x 9 light levels x 2 measurements/plant/light). Observations taken on the same plant are correlated.

For species i, the expected value of NPR, E Y_{ik}, at light intensity X_k is given by:

$$\text{E } Y_{ik} = f(\alpha_i, \beta_i, \gamma_i, X_k).$$

For plant j of species i, the observed NPR Y_{ijk} at light intensity X_k are independent normally distributed random variables with mean

$$\text{E } Y_{ijk} = f(\alpha_{ij}, \beta_{ij}, \gamma_{ij}, X_k)$$

and constant variance.

The investigators want to know if these two species have the same net photosynthesis curve. One way to write this hypothesis is:

$$\text{H0: } \alpha_1 = \alpha_2, \text{ and } \beta_1 = \beta_2, \text{ and } \gamma_1 = \gamma_2$$

(E-1) Write down a reasonable model for the observations. Make sure to define your parameters.

(E-2) Indicate how you would construct a test of the investigator's hypothesis. Describe your test statistic and identify its distribution under the null hypothesis.

(E-3) The investigators will repeat the experiment with two different species. They are considering two possible designs:

A: 10 plants per species, 2 measurements per plant

B: 20 plants per species, 1 measurement per plant

What design would you recommend? Explain your choice.

Part A (A-1) $\mathrm{E}\,Y = \begin{cases} \alpha+\beta X & X < \gamma \\ \alpha+\beta\gamma & X \geq \gamma \end{cases}$

(A-2) Define $X_i^* = \begin{cases} X_i & X_i < \gamma \\ \gamma & X_i \geq \gamma \end{cases}$, then fit the simple linear regression, $Y_i = \alpha + \beta X_i^* + \epsilon_i$.

(A-3) No, because the model is non-linear in γ. Use non-linear least squares (or equivalently maximum likelihood for independent gaussian errors).

Part B (B-1) Define τ as the light compensation point. At $X = \tau$, $0 = \alpha + \beta\tau$. The mle of τ is $-\hat{\alpha}/\hat{\beta} = 1.14$.

(B-2) Using the delta method to approximate the variance of $\hat{\alpha}$ gives:

$$\mathrm{Var}(\hat{\tau}) \approx \frac{\mathrm{Var}(\hat{\alpha})}{\hat{\beta}^2} + \frac{\hat{\alpha}^2\mathrm{Var}(\hat{\beta})}{\hat{\beta}^4} - \frac{2\hat{\alpha}\mathrm{Cov}(\hat{\alpha}\hat{\beta})}{\hat{\beta}^3}$$

$\approx 0.601/0.8523^2 = 0.83$. So, the s.e. of $\hat{\tau} \approx \sqrt{0.83} \approx 0.91$.

Part C (C-1) Unequal variances. Variance of the errors appears to increase with X.

(C-2) Define c.v. of PS = λ. Then, Var(PS) = $\lambda^2(\mathrm{E}\,PS)^2$. Var(Resp) = σ^2 . PS and Respiration are independent, so Var(Net PS) = $\lambda^2(\mathrm{E}\,PS)^2 + \sigma^2 = \lambda^2(\beta\,X)^2 + \sigma^2$, where X is the light intensity.

(C-3) Many possibilities. The data includes replicate observations at each light intensity. One method is to estimate Var(Y) for observations at each light intensity and plot Var(Y) against light intensity. Since the mean Y is a linear function of light intensity up to the light saturation point, the variance model is linear in light intensity up to the light saturation point. This can be assess graphically, or by fitting a regression model to the estimated variances.

(C-4) Many possibilities. Two I expect to see are:
1) Estimated Weighted Least Squares. If Var(Y_i) were known for each observation, α and β could be estimated by weighted least squares. The optimal weights are $w_i = 1/\mathrm{Var}(Y_i)$. Var($Y_i$) are not known, because Var(Y_i) depends on β, λ, X_i and σ^2. However, given β, λ and σ^2 can be estimated by regressing Var(Y) on X_i. Estimates of α, β, λ, and σ^2, can be used to estimate w_i for each observation. Given the estimated w_i, α and β can be estimated by weighted least squares. This procedure can be iterated until convergence.
2) Write a log likelihood function for independent normally distributed observations with $\mathrm{E}\,Y = \alpha + \beta X$ and Var(Y) = $\lambda(\beta X)^2 + \sigma^2$. Find the parameter values that maximize this log likelihood.

Part D (D-1) The desired test is possible. If the two plants have the same light response curve, they have the same intercept and slope. This model is nested in the model with different intercepts and slopes. You can construct an F statistic by comparing the two models. F = (75.76 - 69.93)/(4-2) / (69.93 / (90-4) = 2.915 / 0.813 = 3.58. The 0.95 quantile of an F 2,86 distribution is approximately 3.13. So, $p < 0.05$. Conclusion: Good evidence that the two species do not have the same light response curve. Or, reject the hypothesis that the two species have the same light response curve.

(D-2) This test is not possible here, because the two models are not nested. You could use the difference in residual SS as a test statistic and construct its distribution under H0: no light saturation point using a parametric bootstrap.

Part E (E-1) Any model must allow for variation between plants. So, define:
α_{ij} = intercept for plant j of species i
β_{ij} = slope for plant j of species i
γ_{ij} = light saturation point for plant j of species i
The most straightforward model assumes the plant-plant variation in parameters is a multivariate normal distribution:

$$\begin{bmatrix} \alpha_{ij} \\ \beta_{ij} \\ \gamma_{ij} \end{bmatrix} \sim \text{independent N} \left(\begin{bmatrix} \alpha_i \\ \beta_i \\ \gamma_i \end{bmatrix}, \Sigma \right)$$

Then,

$$Y_{ijk} = f(\alpha_{ij}, \beta_{ij}, \gamma ij, X_k) + N(0, \sigma^2)$$

(E-2) Likelihood ratio test comparing two models: one with three parameters in the mean model (α, β, and γ). This corresponds to the null hypothesis that the two species have the same curve. The second model allows one or more parameters to differ between species. This has six parameters ($\alpha_1, \beta_1, \gamma_1, \alpha_2, \beta_2$, and γ_2) in the mean model. Define L_0 as the log-likelihood for the three parameter model evaluated at the mle's and L_1 as the log-likelihood for the six parameter model evaluated at the mle's. The appropriate test statistic is $-2(L_0 - L_1).Under H0, this has a \chi^2$ distribution with 3 d.f.

(E-3) The error variance, σ^2, is estimated from the variation around each plant's regression line. Σ, the between plant variance-covariance matrix, is estimated from the variability among plants within each species. Both designs provide about the same d.f. to estimate σ^2. Design B provides many more d.f. to estimate Σ, so Design B is preferable.

An experiment was conducted to compare two Drugs A and B, when administered to calves. The response of interest is the concentration of drug in the blood, measured in mg/100ml.

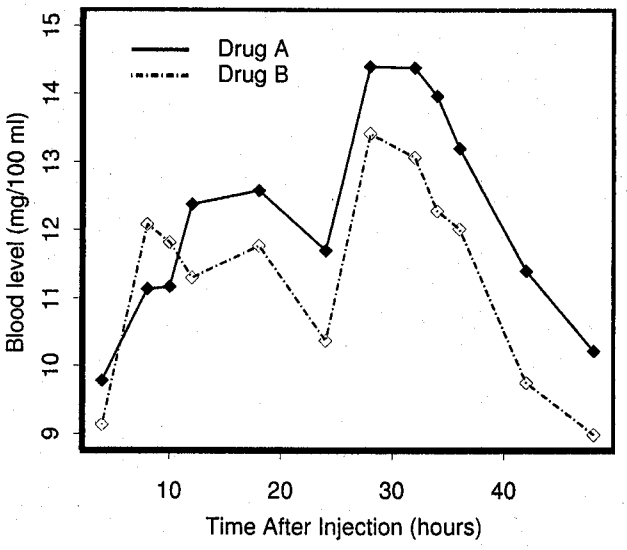
The concentration of drug changes with time. No drug is present in the calf's system at the instant before the drug is administered. Then, the concentration increases as the drug is absorbed and later it decreases as the drug is eliminated. Measurements of drug concentrations were made from blood samples taken at 4, 8, 10, 12, 18, 24, 28, 32, 34, 36, 42, 48 hours after the drug is administered.

The experiment was carried out with 12 calves. Both drugs were used on each calf. After the first drug was administered to a calf, drug concentration in the blood was measured at 4, 8, 10, 12, 18, 24, 28, 32, 34, 36, 42, 48 hours. This was followed by a waiting period to allow the animal to eliminate the drug from its blood and tissue. Then, the other drug was administered and a second set of measurements on drug concentration in the blood were taken at 4, 8, 10, 12, 18, 24, 28, 32, 34, 36, 42, 48 hours after the second drug was administered. Altogether, 24 measurements of drug concentration in the blood were taken on each calf, 12 observations on each drug. Six of the 12 calves were randomly selected to receive Drug A followed by Drug B. The other six received Drug B followed by Drug A.

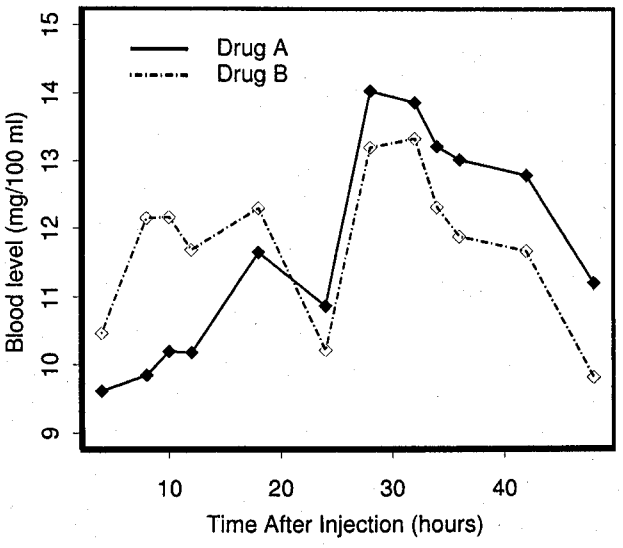
In experiments of this type, large differences among animals can be anticipated. The order in which the drugs are administered may also be a concern for various reasons. One concern is that changes in weather or seasonal factors, such as temperature, can affect responses. For example, absorption of the drug may be suppressed in many animals on very warm days. A second concern is that the first drug administered to a calf could affect the response to the second drug administered to the same calf, even if the researchers wait until the blood concentration of the first drug essentially goes down to zero, because reaction to the first drug that is administered may evoke some unmonitored biological changes in the calf that influence reaction to the second drug that is administered.

A graphical presentation of the averages of the observed drug concentrations is given on the next page. The top graph shows average concentrations of the drugs during the first 48 treatment period. Each point is an average of the observations for 6 calves. Six calves received drug A during the first 48 hour treatment period and the other six calves received drug B. The bottom graph shows drug concentrations during the second 48 hour treatment period. The 6 Drug A calves in the top graph are the 6 Drug B calves in the bottom graph and vice versa.

Mean Concentrations: First Time Period



Mean Concentrations: Second Time Period



The following model was proposed. For the six calves that received Drug A followed by Drug B,

$$Y_{1ijk\ell} = \mu + \alpha_i + \eta_{1j} + \beta_k + \alpha\beta_{ik} + \delta_{1jk} + \tau_\ell + \alpha\tau_{i\ell} + \beta\tau_{k\ell} + \alpha\beta\tau_{ik\ell} + \epsilon_{1jk\ell}$$

where

- η_{1j} is a random effect corresponding to the j-th calf ($j = 1, 2, \ldots, 6$) that received Drug A before Drug B
- α_i is a treatment period effect ($i = 1$ for the first 48 hour period and $i = 2$ for the second 48 hour period
- β_k is a drug effect ($k = 1$ for Drug A and $k = 2$ for Drug B)
- τ_ℓ is a time effect ($\ell = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$ corresponding to 4, 8, 10, 12, 18, 24, 28, 32, 34, 36, 42, 48 hours after the drug is administered)
- δ_{1jk} is a random effect
- $\epsilon_{1jk\ell}$ is a random effect

For the six calves that received Drug B followed by Drug A,

$$Y_{2ijk\ell} = \mu + \alpha_i + \eta_{2j} + \beta_k + \alpha\beta_{ik} + \delta_{2jk} + \tau_\ell + \alpha\tau_{i\ell} + \beta\tau_{k\ell} + \alpha\beta\tau_{ik\ell} + \epsilon_{2jk\ell}$$

where

- η_{2j} is a random effect corresponding to the j-th calf ($j = 1, 2, \ldots, 6$) that received Drug B before Drug A
- α_i is a treatment period effect ($i = 1$ for the first 48 hour period and $i = 2$ for the second 48 hour period)
- β_k is a drug effect ($k = 1$ for Drug A and $k = 2$ for Drug B)
- τ_ℓ is a time effect ($\ell = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$ corresponding to 4, 8, 10, 12, 18, 24, 28, 32, 34, 36, 42, 48 hours after the drug is administered)
- δ_{2jk} is a random effect
- $\epsilon_{2jk\ell}$ is a random effect

It is assumed that $\eta_{1j} \sim NID(0, \sigma^2_\eta)$, $\eta_{2j} \sim NID(0, \sigma^2_\eta)$, $\delta_{1jk} \sim NID(0, \sigma^2_\delta)$, $\delta_{2jk} \sim NID(0, \sigma^2_\delta)$, $\epsilon_{1jk\ell} \sim NID(0, \sigma^2_\epsilon)$, $\epsilon_{2jk\ell} \sim NID(0, \sigma^2_\epsilon)$ and these random effects are stochastically independent of each other.

Sums of squares for an analysis of variance for this model are as follows:

Source	df	Sums of Squares	Mean Squares
Period	1	0.05	0.05
Drugs	1	11.56	11.56
Time	11	438.62	39.87
Period × Drugs	1	15.26	15.26
Period × Time	11	34.24	3.11
Drugs × Time	11	70.71	6.43
Period × Drug × Time	11	9.72	0.88
Calves(Order)	10	110.75	11.08
Calves(Order) × Drugs	10	55.84	5.58
Error	220	390.11	1.77
Corrected Total	287	1136.86	

(a) For this experiment, identify the experimental units and any treatment or blocking factors.

(b) Using $SS_{\text{Drugs}} = 11.56$ in the numerator, construct an appropriate F-statistic. Carefully state the null hypothesis in words (symbols are not sufficient). State your conclusion.

(c) Obtain estimates of the variance components σ^2_η, σ^2_δ, σ^2_ϵ.

(d) This model implies that every observation has the same variance. What does this model imply about correlations among observations taken on the same calf? What does it imply about correlations among observations taken on different calves?

(e) Describe a statistical method for assessing the validity of the covariance structure described in part (d).

(f) The plots on page 2 suggest that the average difference between blood concentrations of the two drugs during the first 24 hour period after the drug is administered is not the same as the mean difference between blood concentrations of the two drugs during the second 24 hour period. To examine this more closely, the researchers constructed the following table of sample means.

	Averages of Concentrations at 4,8,10,12,18, and 24 hours		Averages of Concentrations at 28,32,34,36,42, and 48 hours	
	Drug A	Drug B	Drug A	Drug B
Drug A given first	11.46	11.50	12.93	12.04
Drug B given first	10.39	11.07	13.02	11.59

Obtain standard errors for the following contrasts:

(i) (12.93–12.04) – (13.02–11.59)

(ii) (12.93–11.46) – (12.04–11.50)

(g) Was it important to randomly assign calves to the two possible orders in which the drugs are administered? What are the consequences of doing this random assignment? For example, does it reduce variability, reduce bias, make variances homogeneous, make observations independent, make observations more nearly appear to be normally distributed?

(h) An alternative experiment could have been done using 24 calves, in which 12 calves are randomly assigned to Drug A and the other 12 calves are randomly assigned to Drug B. Drug concentrations in the blood are measured at 4, 8, 10, 12, 18, 24, 28, 32, 34, 36, 42, and 48 hours for each calf, but each calf only uses one drug.

(i) Aside from the increased cost of using 24 calves, what are the potential disadvantages of this design?

(ii) What are the potential advantages of this design? In what circumstances, if any, would it be better than the experiment the researchers actually used?

(i) Suppose the researchers want formulas to describe the trends in blood concentrations of the drugs across time. Consider just the results for the six animals that contributed to the Drug A results shown in the top graph on page 2. The general pattern is that the mean concentration is zero at 0 hours, then it increases and eventually goes back down to zero. One curve that exhibits this pattern is $f(T) = \beta_0 T^{\beta_1} e^{-\beta_2 T}$, for $T > 0$ and $\beta_0 > 0$, $\beta_1 > 0$, $\beta_2 > 0$. Least squares estimation was used to fit this curve to the Drug A results shown in the top graph on page 2, and the estimated coefficients were $\hat{\beta}_0 = 5.371$, $\hat{\beta}_1 = 0.4269$, $\hat{\beta}_2 = 0.0187$.

(i) Describe how you could use a resampling method to make inferences about β_0 and β_1. Be sure to clearly describe how the resampling would be done.

(ii) Would you have any reservations about using resampling methods in this situation? Explain.

(iii) Would you have any reservations about using least squares estimation in this situation? Explain.

(a) The 12 calves are the experimental units. Calves are randomly assigned to one of two orders in which the drugs are administered. Drugs, periods and times within periods are factors corresponding to repeated measurements on the same experimental unit. There are no blocking factors.

(b) $F_{\text{drugs}} = \frac{11.56}{5.58} = 2.07$ on (1,10) d.f. with p-value= 0.18.

The null hypothesis is not rejected at the 0.05 level of significance. The null hypothesis is that the difference in the average blood concentrations for the two drugs is zero, averaging across the six calves in each of the orders in which the drugs were administered, the two orders, and the 12 inspection times. While average blood concentrations were observed to be higher for Drug A, the difference could be a result of random variation among calves.

(c) $\hat{\sigma}_\eta^2 = \frac{\text{MS}_{\text{calves(orders)}} - \text{MS}_{\text{drugs}\times\text{calves(orders)}}}{24} = \frac{11.08-5.58}{24} = 0.23$

$\hat{\sigma}_\delta^2 = \frac{\text{MS}_{\text{calves(orders)}\times\text{drugs}} - \text{MS}_{\text{error}}}{12} = \frac{5.58-1.27}{12} = 0.3175$

$\hat{\sigma}_\epsilon^2 = \text{MS}_{\text{error}} = 1.77$

(d) Any two observations taken on different calves are independent. Any two observations taken on the same calf during the same treatment period (same drug) have correlation $(\sigma_\delta^2 + \sigma_\eta^2)/(\sigma_\epsilon^2 + \sigma_\delta^2 + \sigma_\eta^2)$. Any two observations taken on the same calf during different treatment periods (different drugs) have correlation $\sigma_\eta^2/(\sigma_\epsilon^2 + \sigma_\delta^2 + \sigma_\eta^2)$.

(e) One possibility is a likelihood ratio test against a more complex alternative covariance structure. If the proposed covariance structure is nested in the alternative covariance structure, and the fixed effects are the same for both models, then -2 log(ratio of likelihoods) has approximately a central chi-squared distribution with degrees of freedom equal to the difference in the dimension of the parameter spaces for the two models when the proposed model is correct. You would not be able to test against a general 24×24 covariance matrix because 12 calves are not enough to provide a non-singular estimate of the general 24×24 covariance matrix. A less formal procedure is to compare AIC or BIC values for two difference models. This would allow you to compare the fit of the proposed covariance model against an autoregressive model, or some other model in which it is not nested.

(f) (i) $\sqrt{\frac{1}{18}(\text{MS}_{\text{error}} + \text{MS}_{\text{drugs}\times\text{animals(order)}})} = 0.639$ with approximate degrees of freedom

$$\text{d.f.} = \frac{(\text{MS}_{\text{error}} + \text{MS}_{\text{drugs}\times\text{calves(orders)}})^2}{\frac{(\text{MS}_{\text{error}})^2}{220} + \frac{(\text{MS}_{\text{drugs}\times\text{calves(orders)}})^2}{10}} = 17.3$$

(ii) $\sqrt{\frac{1}{9}\text{MS}_{\text{error}}} = 0.443$ on 220 d.f.

(g) Random allocation of the 12 calves to the two orders in which the drugs can be administered is important to eliminate potential sources of bias due to differences among calves and how they may react differently to different order of treatments. Randomization converts potential sources of biases into random variation. It does not reduce variability in the observed responses Since the calf is the experimental unit that is randomly assigned to the 2 possible orders for administering drugs, this randomization does not necessarily promote homogeneity of variances or reduce correlation among repeated measurements taken on the same calf. It does not promote normality.

(h) (i) Although you have information on twice as many calves, estimates of differences between the overall average concentration of Drug A and Drug B, or the difference between the sample means for the concentrations of Drug A and Drug B at a particular post-injection time, will have much larger variances for this design when there is large variability among animals (strong positive correlation among observations taken on the same animals).

(ii) Advantages of this design are that (1) you do not have to worry about the possibility of carryover effects between treatment periods, (2) you can complete the experiment in 48 hours rather than more than 4 days.

(i) (i) We could employ a bootstrap procedure by taking B samples of size 6 from the six calves that first received Drug A, using simple random sampling with replacement, and obtain least squares estimation of the parameters in the formula for the curve for each sample. Results from the B samples could be used to estimate the covariance matrix for the parameter estimates or construct confidence intervals or regions.

(ii) The original set of 6 calves may not represent a random sample from the population of calves that could have been used in this experiment. Even if the calves were randomly selected from some larger population, the sample cdf for such a small sample may deviate substantially from the population cdf.

(iii) If the proposed curve accurately describes the trend in the mean blood concentration of the drug, least squares estimation would yield consistent estimates of the parameters. In that sense, least squares estimation is not bad. Since the observations are correlated across time, least squares estimators are not efficient. Better estimators could possibly be obtained ifwe could include accurate information about the correlations in the estimating equations. This would be difficult with information from only six calves.

In a study of the effects of an experimental chemical compound on the general reproductive performance in laboratory rats, the outcome of primary interest was taken to be the weight of rat pups. The experimental compound was administered through diet (food) at three levels, a control, a low dose, and a high dose. Thirty female mice (dams) were randomly allocated into the three treatment groups in a way that resulted in 10 dams receiving the control treatment, 10 dams receiving the low dose treatment, and 10 dams receiving the high dose treatment. In the high dose group, however, one female did not conceive, one cannibalized her litter, and one delivered only a single still-born pup, so that the high dose treatment produced data from only 7 litters. Litter size ranged from 2 to 18, and recorded data for each pup included sex (male or female) and weight. There were a total of 27 litters containing 321 records for individual pups.

The overall objective in this study was to investigate the effect of the experimental compound on pup weight, within the context of weight as an indicator of general reproductive success. A number of tables of summary statistics and exploratory data plots are presented at the end of this question (pages 10-18) to assist you in formulating your answers to the questions posed. Some of these may be meaningful for the problem under discussion and some may not be meaningful. In these summaries, treatment 1 is the control group, treatment 2 is the low dose group, and treatment 3 is the high dose group. Dams are labeled consecutively from 1 to 27.

1. Examine the plots and tables of summary statistics presented at the end of this question. *Do not ignore the fact that you have done so in answering certain other portions of the questions that follow.*

   1A. List the factors or variables that should be included in an initial model for this problem. Indicate what led you to decide that each factor or variable you list should be included (consider whether there are any indications in the summary tables or exploratory data plots that certain factors or variables are likely to be important in the analysis).

   1B. In particular, does figure 7 provide useful information concerning whether an assumption of normality appears reasonable for error terms associated with any linear model that might be used for analysis of these data? Why or why not?

2. A one way analysis of variance was fit to these data with the following results:

Source	df	SS	MS	F	p
Trt	2	13.95	6.98	18.49	< 0.00001
Error	318	119.93	0.38		

Trt Contrast	Point Estimate	95% Interval
1 - 2	0.407	(0.226, 0.587)
1 - 3	0.456	(0.235, 0.677)
2 - 3	0.049	(-0.172, 0.271)

2A. Write the model that corresponds to the overall ANOVA table given above. Be certain to define any quantities used in your notation.

2B. From a statistical viewpoint, what would be the *primary* objection to using this model for the study under consideration?

3. A similar model to that of question 2 was fit to the litter means, with the following results:

Source	df	SS	MS	F	p
Trt	2	1.005	0.5024	1.50	0.2433
Error	24	8.038	0.3349		

Trt Contrast	Point Estimate	95% Interval
1 - 2	0.409	(-0.237, 1.060)
1 - 3	0.384	(-0.328, 1.102)
2 - 3	-0.025	(-0.738, 0.687)

3A. Given that the point estimates of treatment contrasts are similar to those for the model of question 2, why is there such a dramatic difference in the conclusions that would be reached from this analysis?

3B. Could the systematic component of this model (i.e., the mean structure) be extended to incorporate the factors or variables you identified as of potential importance in question 1? Why or why not?

3C. If the model of question 2 is appropriate, does it surprise you that the estimate of error variance from this analysis (0.33) is so similar to that from the analysis of question 2 (0.38)? Why or why not?

4. Two other linear models might be suggested for this problem. In the first model, take Z_i to represent the mean weight of pups from litter i; $i = 1, \ldots, 27$, and let

$$Z_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \sigma \epsilon_i; \quad \epsilon_i \sim \ iid N(0,1), \tag{1}$$

$\boldsymbol{x}_i^T$ is a 1×4 vector with terms for an intercept, two contrasts for treatment, and the numeric covariate litter size. Treatment contrasts were coded as two indicator variables, the first having the value 1 if treatment level was 2 (low dose) and the second having the value 1 if treatment level was 3 (high dose). The control treatment was absorbed into the intercept term.

In the second model, take $Y_{i,j}$ to represent the weight of pup j in litter i; $j = 1, \ldots, n_i$ and $i = 1, \ldots, 27$, and let

$$Y_{i,j} = \boldsymbol{x}_{i,j}^T \boldsymbol{\beta} + \delta_i + \sigma \epsilon_{i,j}; \quad \epsilon_{i,j} \sim iid\ N(0,1), \tag{2}$$

where $\boldsymbol{x}_{i,j}^T$ is the same 1×4 vector as for model (1), replicated for each pup (index j) within the same level of litter i, and $\delta_i \sim N(0, \sigma_d^2)$ is a random effect for litter (dam) i; $i = 1, \ldots, 27$.

Finally, the most elaborate model fit to these data was of the same form as model (2) but with an extra contrast added to the covariate vector $\boldsymbol{x}_{i,j}^T$ for sex (in the form of an indicator variable for female). Thus, $\boldsymbol{x}_{i,j}^T$ now becomes a 1×5 vector. We will call this the extended version of model (2).

A fit of model (1) resulted in

Term	Estimate	Std. Error	t	p
Intercept	8.074	0.2312	34.91	< 0.0001
Trt 2	-0.446	0.1400	-3.19	0.0041
Trt 3	-0.861	0.1162	-5.18	< 0.0001
Litter Size	-0.124	0.0161	-7.68	< 0.0001

The estimated value of σ^2 in this model was $\hat{\sigma}^2 = 0.0979$.

A fit of model (2) resulted in

Term	Estimate	Std. Error	t	p
Intercept	8.068	0.2644	30.51	< 0.0001
Trt 2	-0.450	0.1420	-3.17	0.0043
Trt 3	-0.885	0.1740	-5.09	< 0.0001
Litter Size	-0.123	0.0184	-6.68	< 0.0001

The estimated value of σ^2 in this model was $\hat{\sigma}^2 = 0.1959$, and the estimated value of σ_d^2 was $\hat{\sigma}_d^2 = 0.0828$.

A fit of the extended version of model (2) resulted in:

Term	Estimate	Std. Error	t	p
Intercept	8.305	0.2755	30.147	< 0.0001
Trt 2	-0.406	0.1511	-2.68	0.0133
Trt 3	-0.871	0.1834	-4.75	0.0001
Sex	-0.361	0.0476	-7.58	< 0.0001
Litter	-0.129	0.0191	-6.79	< 0.0001

The estimated value of σ^2 in this model was $\hat{\sigma}^2 = 0.1632$, and the estimated value of σ_d^2 was $\hat{\sigma}_d^2 = 0.0986$.

4A. Give a brief argument that model (1) is appropriate for this analysis and would be preferable to the use of model (2) or its extended version.

4B. Give a modification of model (1) that would *a priori* reflect the structure of the data more accurately than does expression (1).

4C. Give a brief argument that model (2) is appropriate for this analysis and would be preferable to the use of model (1).

4D. Give a brief argument that the extended version of model (2) is appropriate for this analysis and would be preferable to either model (1) or the original model (2).

4E. Are estimates of the variance components for models (1) and (2) or its extended version more compatible than those of the two ANOVA models in questions 2 and 3? Explain.

5. Models with random components, such as model (2) or its extended version from question 4 may be written as,

$$Y_{i,j} = \boldsymbol{x}_{i,j}^T \boldsymbol{\beta} + \delta_i + \epsilon_{i,j}, \tag{3}$$

where

$$\begin{aligned} \epsilon_{i,j} &\sim iid\ N(0, \sigma^2), \\ \delta_i &\sim iid\ N(0, \tau^2), \end{aligned}$$

and $\{\epsilon_{i,j} : i = 1, \ldots, K;\ j = 1, \ldots, n_i\}$ are assumed independent of $\{\delta_i : i = 1, \ldots, K\}$.

5A. Under model (3), what are the expected value and variance of $Y_{i,j}$?

5B. Under model (3), what are the covariances of $Y_{i,j}$ and $Y_{i,m}$, $m \neq j$? Of $Y_{i,j}$ and $Y_{q,m}$, $q \neq i$, $m \neq j$?

5C. In models that follow the general form of (3) we may think of random variables $Y_{i,j}$ as occurring in groups or "clusters", and compare the between cluster variance to the within cluster variance, typcially by examining the relative magnitudes of the variances σ^2 and τ^2. Several quantities can be used to express the effects of within cluster and between cluster variability relative to one another. For an individual random variable $Y_{i,j}$ the effect of τ^2 relative to σ^2 may be quantified as τ^2/σ^2, or the effect of τ^2 may be quantified as relative to the variance of $Y_{i,j}$ that you derived in question 5A. Consider now the *total* variability within a given cluster (level of i). Let the sum of residual quantities within a cluster i be,

$$\sum_{j=1}^{n_i} \left(Y_{i,j} - \boldsymbol{x}_{i,j}^T \boldsymbol{\beta} \right).$$

Use this quantity to determine the relative contributions of the two error variances σ^2 and τ^2 to the total variability within a given cluster (level of i).

5D. What does your result from question 5C indicate about the effect of the random component variance τ^2 relative to the cluster size n_i?

5E. Is the effect identified in question 5D important in this scientific problem? That is, would we expect this effect to be realized in studies of the type considered here? Why or why not?

6. For this question do not necessarily restrict your attention to linear models.

6A. Define possible observable random variables at the level of a litter in the experiment.

6B. Define possible observable random variables at the level of an individual rat pup in the experiment.

6C. Using some combination of the random variables defined in parts 6A and 6B, construct an alternative model for this experiment. Do you think your model might have possible advantages over those given in questions 2-5 (don't forget you have exploratory plots and tables)? How might you go about estimation and inference with your model (no derivations, just indication of what methods you might use)?

Note: Question 6C is somewhat open-ended, but it is NOT worth more than any of the other questions posed previously. Do not spend all your time writing an extensive answer to 6C while failing to complete the previous questions.

Table of Contents
Summary Tables and Exploratory Plots

1. Table 1. Summary Statistics for Treatment Groups, page 10.
2. Table 2. Litter Sizes for Treatment Groups, page 10.
3. Table 3. Summary Statistics for Individual Litters, page 11.
4. Figure 1. Boxplots of rat pup weights for treatment groups, page 12.
5. Figure 2. Boxplots of rat pup weights for treatment groups, page 12.
6. Figure 3. Mean weights over all pups in litters versus litter size, page 13.
7. Figure 4. Proportion of male pups in litters versus litter size, page 13.
8. Figure 5. Mean weight of male and female pups versus litter size, taken over all treatments, page 14.
9. Figure 6. Meand weight of male and female pups versus the number of that sex in litters, taken over all treatments, page 14.
10. Figure 7. Histogram of 321 rat pup weights over all litters and treatments, page 15.
11. Figure 8. Histogram of 27 litter sizes over all treatments, page 15.
12. Figure 9. Boxplots of proportion of male pups in litters by treatment group, page 16.
13. Figure 10. Boxplots of litter sizes by treatment group, page 16.
14. Figure 11. Scatterplot of mean male and female weights in litters, page 17.
15. Figure 12. Ratios of mean male to mean female pup weights versus litter size, page 17.
16. Figure 13. Standard deviations versus means of male and female pups within litters, page 18.
17. Figure 14. Standard deviations of male and female pup weights within litters versus the number of that sex, taken over all treatments, page 18.

Table 1. Summary Statistics for Treatment Groups

	Number of Pups		Proportion of	Average Weight		Standard Deviation	
Trt	Male	Female	Male Pups	Male	Female	Male	Female
1	76	54	0.5846	6.487	6.116	0.7451	0.6851
2	62	66	0.4843	6.019	5.843	0.3796	0.4493
3	33	31	0.5156	5.918	5.833	0.6909	0.6009

Table 2. Litter Sizes for Treatment Groups

	Litter Size	
Trt	Mean	Variance
1	13.0	16.67
2	12.7	17.34
3	9.14	12.14
All	11.89	17.33

Table 3. Summary Statistics for Individual Litters (Dams)

		Number of Pups		Proportion of	Average Weight		Standard Deviation	
Trt	Dam	Male	Female	Male Pups	Male	Female	Male	Female
1	1	8	4	0.6666	6.945	6.645	0.4326	0.2830
1	2	9	5	0.6428	6.485	5.956	0.2699	0.0920
1	3	2	2	0.5000	7.290	7.420	0.2969	0.2121
1	4	8	6	0.5714	6.495	6.015	0.2747	0.3380
1	5	8	5	0.6153	7.132	6.490	0.6767	0.9479
1	6	6	3	0.6666	7.528	5.840	0.8388	1.9012
1	7	9	9	0.5000	6.351	6.118	0.2176	0.2337
1	8	8	9	0.4705	6.061	5.745	0.4278	0.5339
1	9	12	4	0.7500	5.390	5.320	0.3241	0.1283
1	10	6	7	0.4615	6.666	6.421	0.3280	0.2364
2	11	10	6	0.6250	5.836	5.540	0.2261	0.1206
2	12	0	2	0.0000	NA	7.310	NA	0.5939
2	13	9	4	0.6923	5.858	5.627	0.2704	0.4385
2	14	9	6	0.6000	5.928	5.656	0.2101	0.1889
2	15	7	6	0.5384	6.195	5.936	0.1684	0.2170
2	16	3	10	0.2307	6.143	5.967	0.2802	0.2620
2	17	7	7	0.5000	5.544	5.357	0.2442	0.2608
2	18	3	12	0.2000	6.916	5.951	0.1890	0.4323
2	19	4	6	0.4000	6.320	6.143	0.4172	0.1966
2	20	10	6	0.6250	6.213	5.760	0.2296	0.3415
3	21	11	3	0.7857	5.323	4.946	0.2224	0.4072
3	22	5	5	0.5000	5.508	5.306	0.3063	0.1402
3	23	1	2	0.3333	7.700	7.005	NA	0.9545
3	24	3	9	0.2500	6.103	5.763	0.3146	0.3536
3	25	5	3	0.6250	6.690	6.353	0.5248	0.2657
3	26	2	6	0.2500	6.575	6.185	0.6010	0.2241
3	27	6	3	0.6666	6.100	5.803	0.3732	0.1096

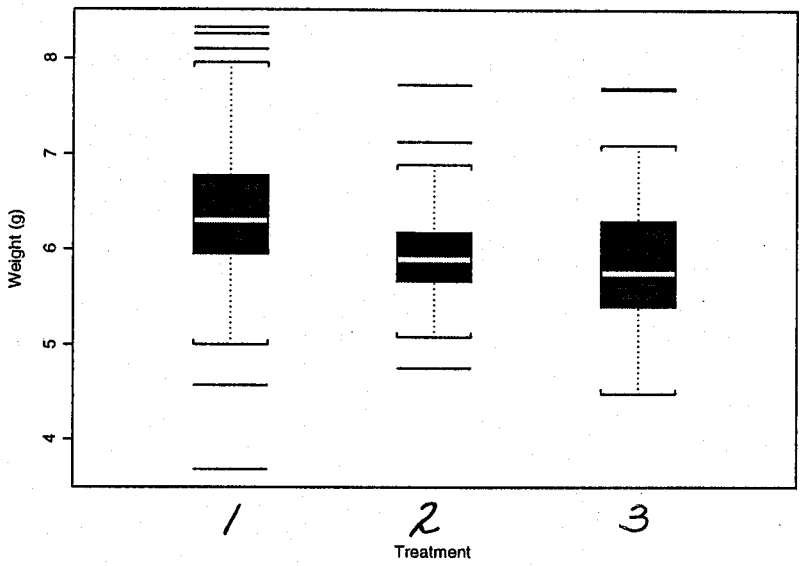


Figure 1.Boxplots of rat pup weights for treatment groups.

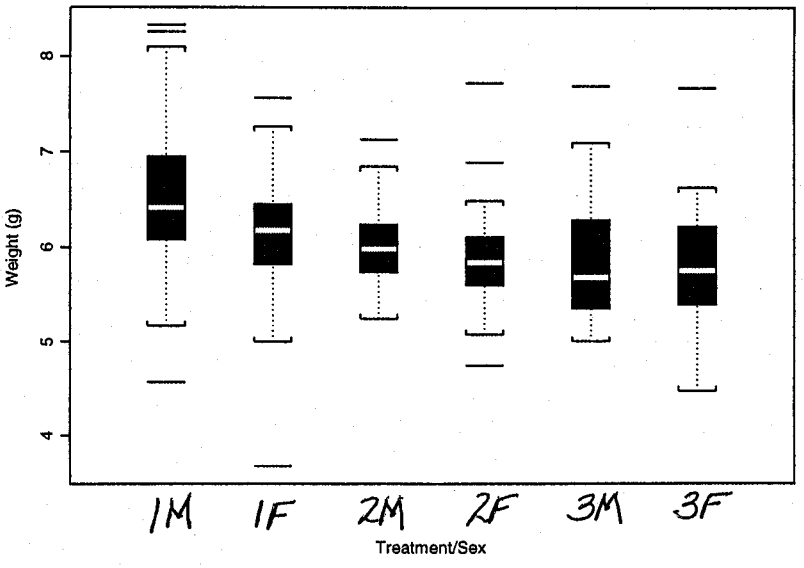


Figure 2.Boxplots of rat pup weights for male and female pups from treatment groups.

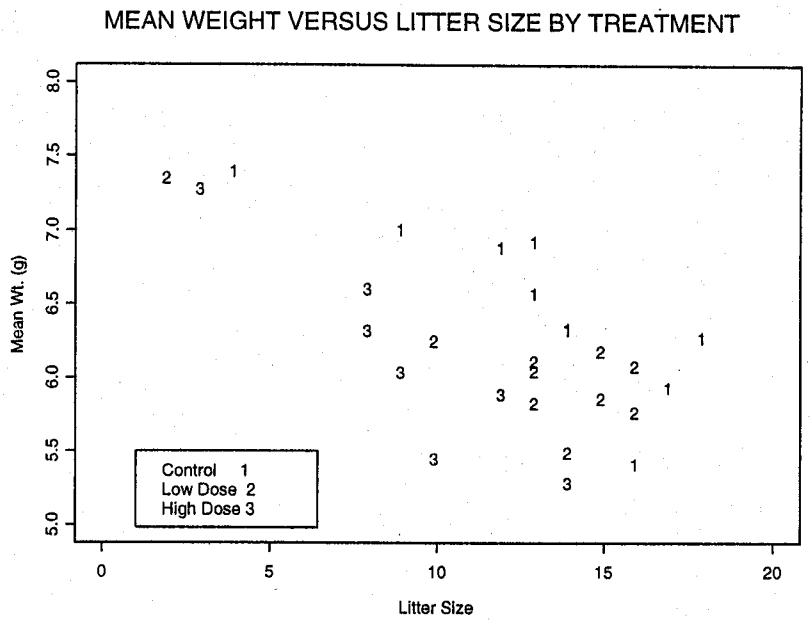


Figure 3.Mean weights over all pups in litters versus litter size.

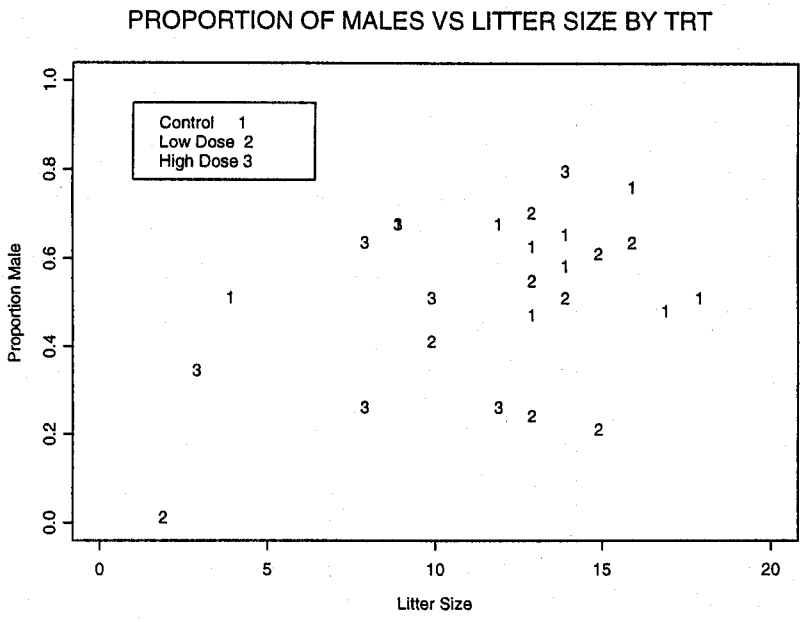


Figure 4.Proportion of male pups in litters versus litter size.

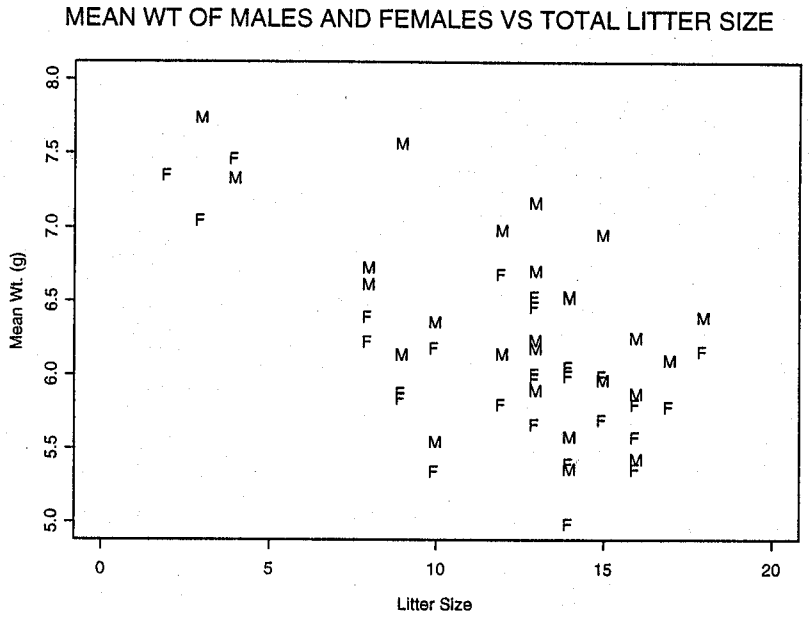


Figure 5.Mean weight of male and female pups versus litter size, taken over all treatments.

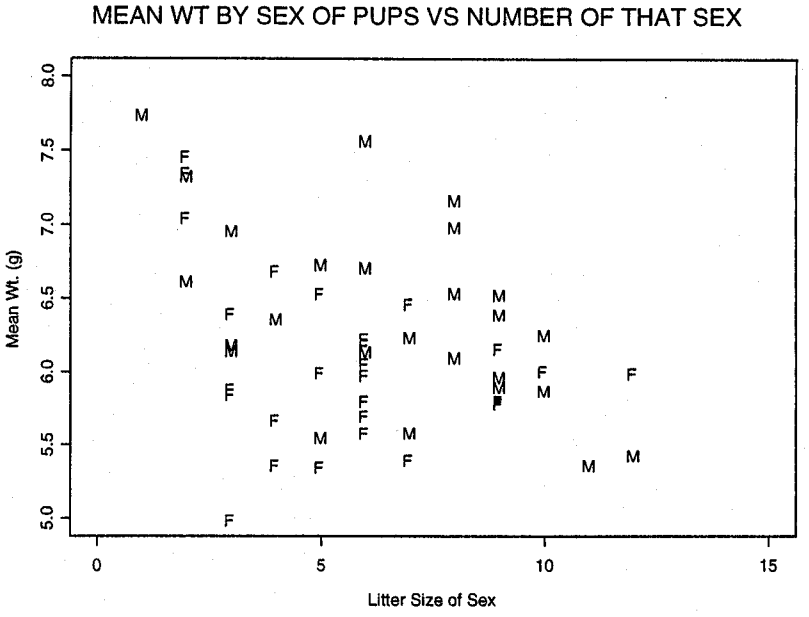


Figure 6.Mean weight of male and female pups versus the number of that sex in litters taken over all treatments.

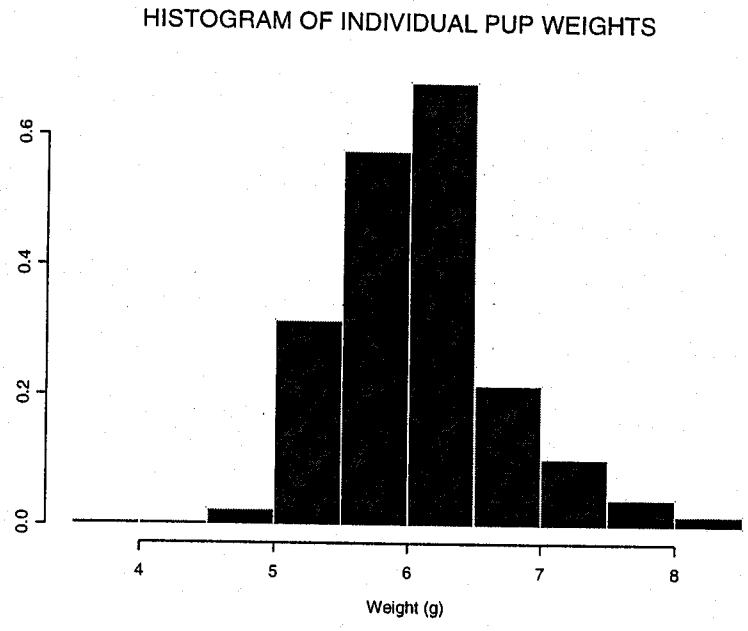


Figure 7.Histogram of 321 rat pup weights over all litters and treatments.

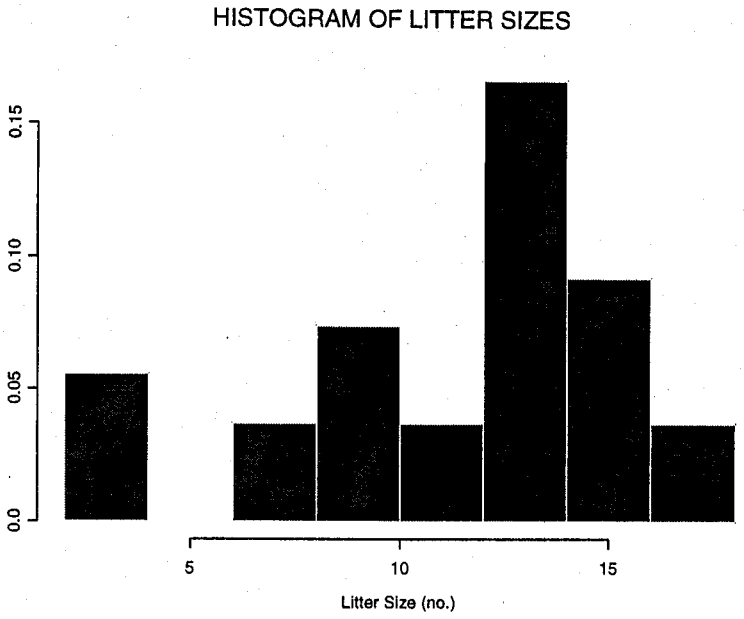


Figure 8.Histogram of 27 litters over all treatments.

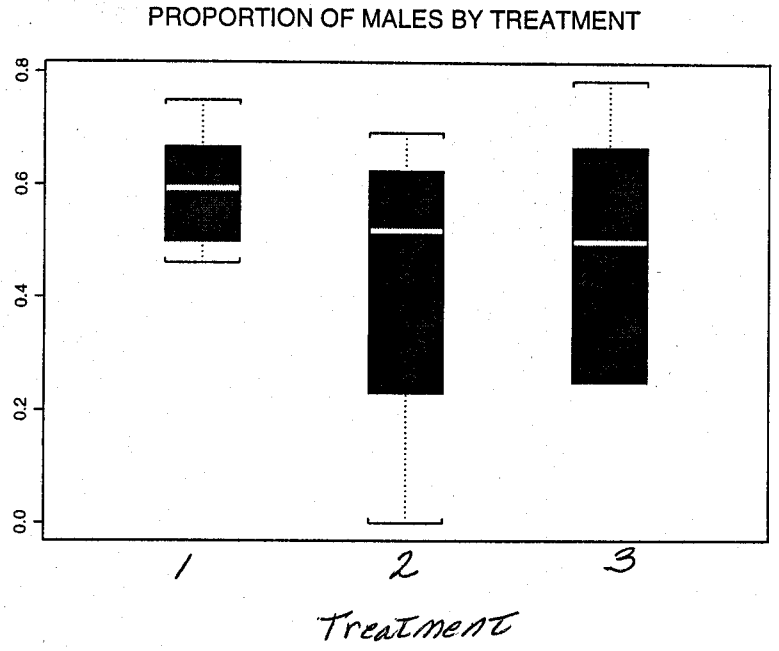


Figure 9.Boxplots of proportion of male pups in litters by treatment group.

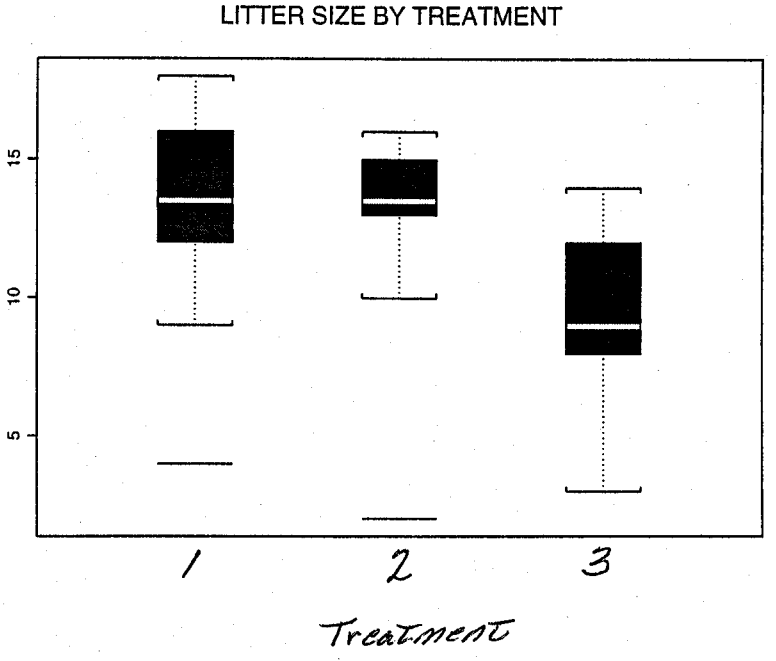


Figure 10.Boxplots of litter sizes by treatment group.

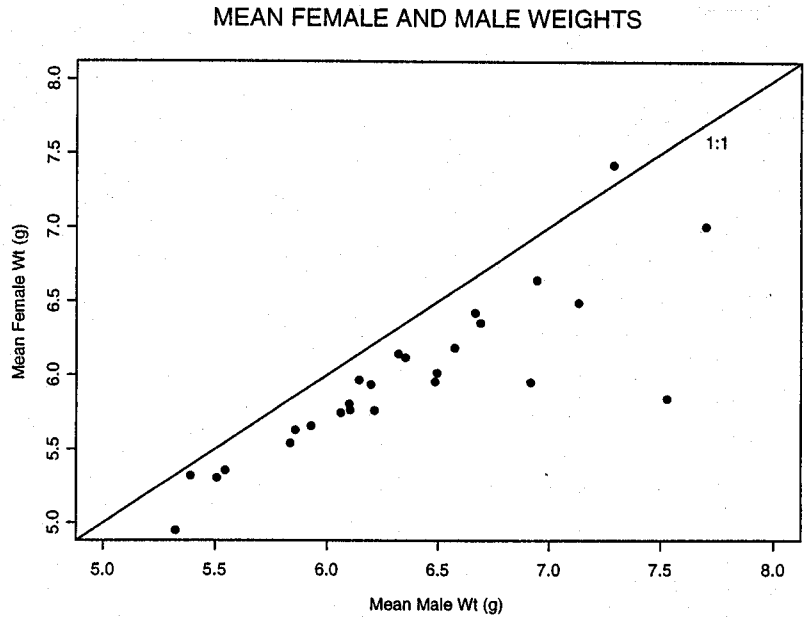


Figure 11.Scatterplot of mean male and female weights in litters (i.e., each point is one litter) over all treatments.

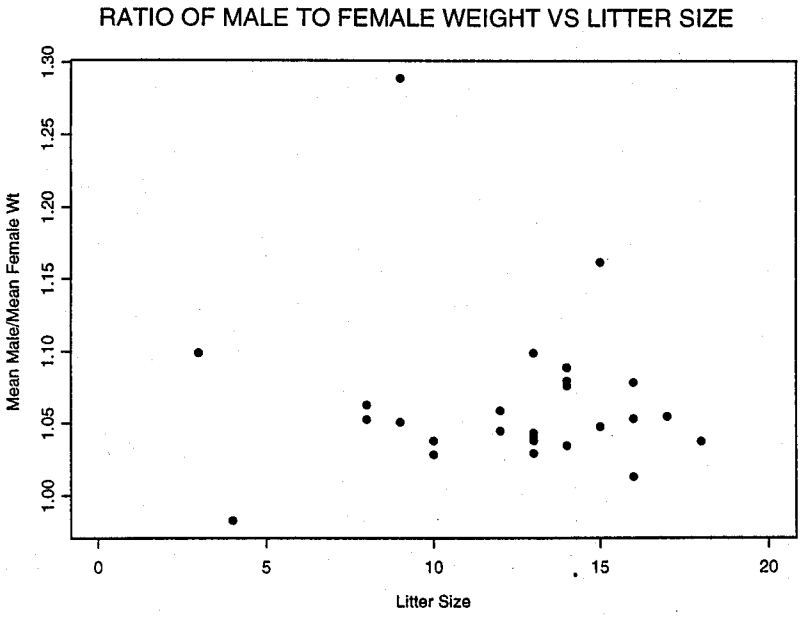


Figure 12.Ratios of mean male to mean female pup weights versus litter size.

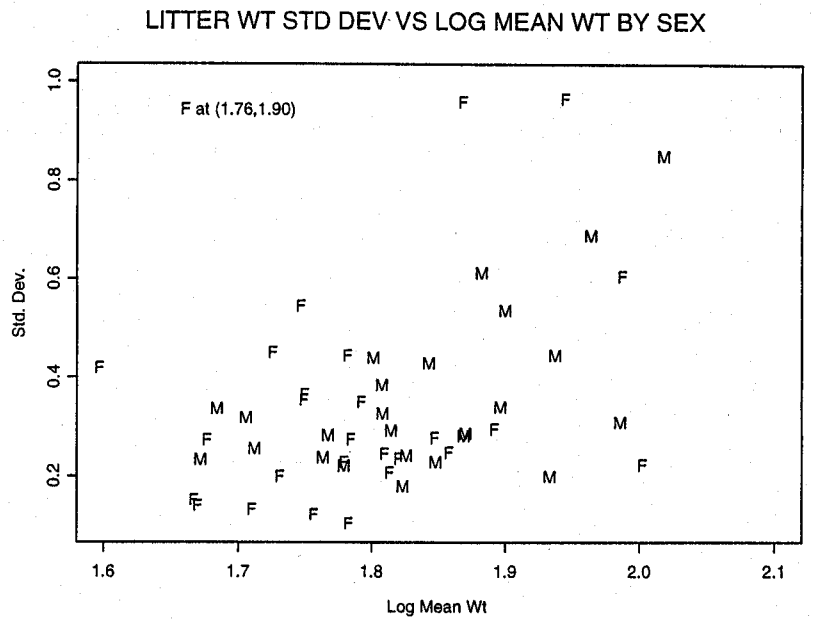


Figure 13.Standard deviations vs. means of male and female pups within litters.

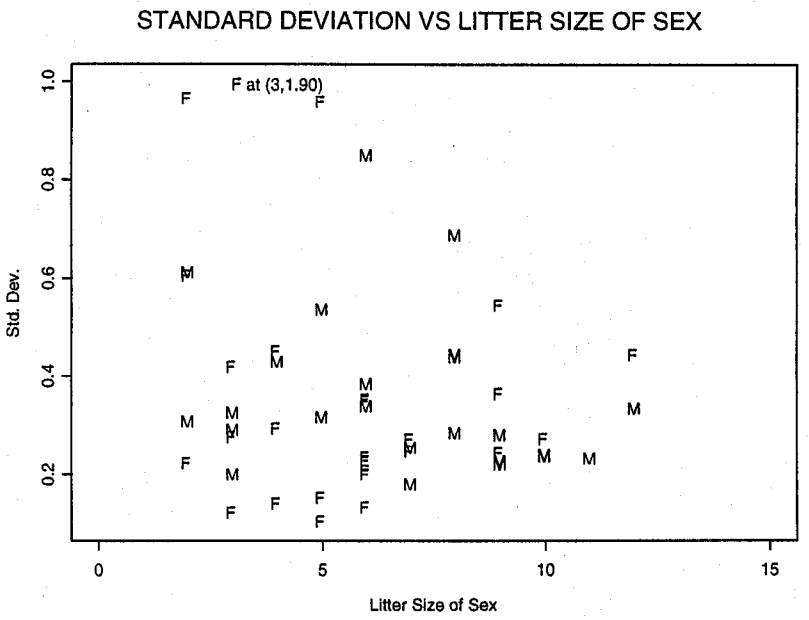


Figure 14.Standard deviations of male and female pups within litters versus the number of that sex in litters. Taken over all treatments.

1A. The following points are relevant:

1. From the boxplot of Figure 1 it is not clear that treatment has an overall effect. Nevertheless, the objective of the study was to determine the possible effect of treatment so that a treatment effect must necessarily be included in any model considered. A prior opinion on this factor would be that it is probably not important, unless there are other factors involved that mask treatment effect unless they are adjusted for. A stronger indication that this may be the case is given in Figure 3 in which it appears that mean weights for treatment 1 lie above those of treatment 2 which in turn lie above those of treatment 3.

2. Litter size appears to be relatively important factor in this study. This is indicated by Figure 3 as discussed above, Figure 5 which shows the same data with each value from Figure 3 split into two components. Interestingly, litter size by treatment does not show a strong association with treatment in Figure 10, although it does appear that litter sizes were smaller for the high dose treatment (treatment 3). Given that the effect of treatment appears clear only when litter size is accounted for (i.e., Figure 3 versus Figure 1), it should be anticipated that litter size will be an important factor in the analysis.

3. That sex of rat pups is related to individual weight is illustrated clearly in Figure 11, with mean weight of males in litters nearly always greater than mean weight of females. That the degree to which (mean) male weight was greater than that of (mean) female weight was relatively constant across all litters in the study and, in particular, was not related to litter size, is illustrated in Figure 12. While there is clear indication that male rat pups are, in general, heavier than female pups, it is not clear that this factor is related to treatment (Figure 4, Figure 9). It may well be important in the analysis since there appears greater variability in the proportion of male pups in litters from treatment 2 and treatment 3 groups than for the treatment 1 group.

1B. The histogram of Figure 7 presents a picture of the marginal distribution of response variables, which provides no useful information relative to a random component for any model that includes a varying mean structure among groups of experimental or sampling units in the study. While this histogram certainly does not contradict an assumption of normality, it does not support such an assumption either. What would be important is the empirical distribution of appropriately defined residual quantities corresponding to some modeled mean structure.

2A. Let $Y_{i,j}$ correspond to the weight of pup j from litter i, $j = 1, \ldots, n_i$, $i = 1, \ldots, 27$. A model for the overall ANOVA table would be

$$Y_{i,j} = \sum_{t=1}^{3} \{I(Y_{i,j} \in \text{trt group } t)\mu_t\} + \sigma\epsilon_{i,j}; \qquad \epsilon_{i,j} \sim iid\ N(0,1),$$

where $\{\mu_t : t = 1, 2, 3\}$ are parameters to be estimated and σ^2 is the common variance of additive error terms.

2B. The *primary* objection to this model would be the assumption that the $\epsilon_{i,j}$ are independent. That is, experimental units in this study were female rats, which correspond to litters. Thus, treatments were not independently applied to individual pups within litters. In fact, the litter sizes $\{n_i : i = 1, \ldots, 27\}$ are not under the control of the investigator. If there is any association between litter size and treatment, due to either treatment effects or uncontrolled randomness, this model will provide inefficient and perhaps unreliable estimates of treatment effects.

3A. The difference in conclusions from this analysis and that of question 2 is entirely a matter of assumed sample size. Under the assumption that the model of question 2 holds, the results are highly significant due mainly to the number of "replicates" that exist under that model.

3B. This model could be extended in mean structure to include any factors or variables observed at the litter level. In particular, this would include litter size but not sex of individual rat pups.

3C. If the model of question 2 were appropriate it would be highly surprising that the estimated error variance from a model applied to aggregated data (i.e., litter means) would have the same value. Under the model of question 2 the litter means, taken as the fundamental response variables in this analysis, should have variance equal to σ^2/n_i. While the litter sizes n_i in this study were quite variable (range 2 to 18, see Table 3) one would still expect to see the effect of litter sizes on error variance of means.

4A. Model (1) is appropriate as it uses experimental units of litters (dams) as the fundamental unit of response. The error variance in this model is entirely due to variability among litters which appears from consideration of the two ANOVA models in question 2 to be the primary source of variation among responses. Since interest in this study is primarily on assessment of the treatment effects, there is little gain from considering more detailed models for the error structure. Inferences from this model apply to the population of litters among lab rats similar to those used in the study.

4B. Since the response variables in model (1) are taken to be litter means, and since the size of litters varies, a more appropriate structure would seem to be,

$$Z_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \frac{\sigma}{\sqrt{n}}\epsilon_i; \quad \epsilon_i \sim iid\ N(0,1),$$

where $\boldsymbol{x}_i^T$ is as defined in the question and n_i is the size of the ith litter. This is because, as written in the question text, model (1) takes the variance of the litter means (the Z_i) to be constant, although it is known *a priori* that the litter sizes differ.

4C. Model (2) is appropriate since it results in individual pup weights being independent among litters but correlated within litters. This allows use of unaggregated responses at the level of sampling units (i.e., pups). Through use of this model we can assess not only among litter variance (the estimate of which is nearly identical to that of model (1)) but also among individual pups within litters. Thus, the model is a more accurate reflection of the combined experimental and observational processes of the study. Inferences from this model apply to the population of pups, which have the same expected values as litters, but also allows inferences on the sources of variation present.

4D. The extended version of model (2) is appropriate for the same reasons given for the original model (2). In addition, this extended model incorporates the effect of pup sex on the outcome of interest (i.e., pup weight). The exploratory analyses of question 1 clearly indicated a systematic effect of this factor and its incorporation into the model allows a more detailed description of the overall situation. Taking sex of pup into account reduces the relative contribution of within litter variance. For model (2) the relative magnitude of between and within litter variances was $\hat{\sigma}_d^2/\hat{\sigma}^2 = 0.4227$. Taking this source of variation into account in the extended version of model (2) increases the random between litter contribution to $\hat{\sigma}_d^2/\hat{\sigma}^2 = 0.6042$.

4E. Yes, the estimated variances between model (1) and model (2) appear to be more compatible than those for the two ANOVA models. The variance σ^2 in model (1) reflects variability between litters, since the response variables are litter means. This should correspond to the random effect variance σ_d^2 in model (2) or its extended version, and the estimated values of $\hat{\sigma}^2 = 0.0979$ from model (1) compares quite favorably with the estimate $\hat{\sigma}_d^2 = 0.0828$ from model (2) or $\hat{\sigma}_d^2 = 0.0986$ from the extended version of model (2).

5A.

$$\begin{aligned} E\{Y_{i,j}\} &= E\left\{\boldsymbol{x}_{i,j}^T\beta + \delta_i + \epsilon_{i,j}\right\} \\ &= \boldsymbol{x}_{i,j}^T\beta. \end{aligned}$$

$$\begin{aligned} var\{Y_{i,j}\} &= var\left\{\boldsymbol{x}_{i,j}^T\beta + \delta_i + \epsilon_{i,j}\right\} \\ &= var\{\delta_i\} + var\{\epsilon_{i,j}\} \\ &= \tau^2 + \sigma^2. \end{aligned}$$

5B.

$$\begin{aligned} cov\{Y_{i,j}, Y_{i,m}\} &= \frac{1}{2}\left(var\{Y_{i,j} + Y_{i,m}\} - var\{Y_{i,j}\} - var\{Y_{i,m}\}\right) \\ &= \frac{1}{2}\left(var\{\boldsymbol{x}_{i,j}^T\beta + \boldsymbol{x}_{i,m}^T\beta + 2\delta_i + \epsilon_{i,j} + \epsilon_{i,m}\} - var\{Y_{i,j}\} - var\{Y_{i,m}\}\right) \\ &= \frac{1}{2}\left(4\tau^2 - 2(\sigma^2 + \tau^2) - \sigma^2 - \sigma^2\right) \\ &= \tau^2. \end{aligned}$$

Alternatively, this may be done directly as,

$$cov\{Y_{i,j}, Y_{i,m}\} = E\{Y_{i,j}Y_{i,m}\} - E\{Y_{i,j}\}E\{Y_{i,m}\}$$

which leads to the same result. For variables from different litters,

$$cov\{Y_{i,j}, Y_{q,m}\} = 0.$$

5C. We can accomplish the objective of this question by looking at the variance of the sum of residuals within a given group,

$$\begin{aligned} var\left\{\sum_{j=1}^{n_i}\left(Y_{i,j} - \boldsymbol{x}_{i,j}^T\beta\right)\right\} &= var\left\{\sum_{j=1}^{n_i}(\delta_i + \epsilon_{i,j})\right\} \\ &= var\left\{n_i\delta_i + \sum_{j=1}^{n_i}\epsilon_{i,j}\right\} \\ &= n_i^2\tau^2 + n_i\sigma^2. \end{aligned}$$

For the sum of variances within a given level of the index i the contribution of σ^2 and tau^2 can be written as

$$\frac{n_i\sigma^2}{n_i^2\tau^2} = \frac{1}{n_i\tau^2/\sigma^2}.$$

5D. The implication of the result in 5C is that even fairly small random component variance τ^2 can have a great effect on the total variation within a cluster for large cluster sizes. For example, suppose that the random component variance τ^2 is only $1/10$ of the within cluster variance σ^2. For a cluster of size 10 this gives a sum of residual variances equal to $20\sigma^2$. But when cluster size is increased to 100 the result is $1100\sigma^2$. In other words, if $\tau^2 = 0$ a ten times increase in cluster size would lead to a ten times increase in total variation. But, with $\tau^2 = 0.1\sigma^2$, a ten times increase in cluster size leads to a 55 times increase in total variation.

5E. No, this effect would not be expected to be realized in this situation. Clusters in this study correspond to litters. Litter size in lab rats is physiologically constrained.

6A. Observable random variables at the level of a litter would include:

(a) Litter size, a count

(b) Number of male pups, a count

(c) Number of female pups, a count

6B. Observable random variables at the level of rat pups would include:

(a) Weight, a continuous variable on the positive line

(b) Sex, a binary variable

6C. There are any number of possibilities here, one of which would be to assign (within a single treatment group) the following conditional and marginal distributions to pup weight, pup sex, and litter size.
Let $Y_{i,j}$ represent weight of pup j in litter i, $j = 1, \ldots, n_i$, $i = 1, \ldots, K$, $Y_{i,j} \in (0, \infty)$.

Let $S_{i,j}$ represent sex of pup j in litter i, $j = 1, \ldots, n_i$, $i = 1, \ldots, K$, $S_{i,j} = 1$ if male, $S_{i,j} = 0$ if female.

Let n_i represent size of the ith litter, $i = 1 \ldots, K$, $n_i \in \{0, 1, \ldots\}$.

Let the conditional distribution of $Y_{i,j}$ given $S_{i,j}$ and n_i be

$$[Y_{i,j}|S_{i,j}, n_i] = N(\mu_{n,s}, \sigma_s^2)$$

where,

$$\mu_{s,n} = \beta_0 + \beta_1 n_i + \beta_2 S_{i,j}.$$

Here, we build into the conditional expectation of pup weight much the same structure as included in the extended version of model (2) in this question. A difference here is that we might allow the variance to differ for male and female pups. There is some suggestion that this might be appropriate from Figures 13 and 14. Further, model the sex of pups as

$$[S_{i,j}|\theta_i] = \text{ Binary } (\theta_i)$$

$$[\theta_i] = \text{ Beta } (\alpha, \beta).$$

Here, we allow the probability of a male pup to vary among litters based on the evidence of Figure 4. Finally, we might model litter size as,

$$[n_i|\lambda_i] = \text{ Po } (\lambda_i)$$

$$[\lambda_i] = \text{ Gamma } (\alpha^*, \beta^*).$$

Similar to the modeling of $S_{i,j}$ we take the marginal distribution of n_i to be a mixture distribution, which is suggested by the values of Table 2 demonstrating that mean litter size (both within and over treatments) is only about 75% of the variance. If we wished to conduct a Bayesian analysis of this model we would further need prior distributions on β, σ_m^2 and σ_f^2, where m and f denote data model variances for males and females, respectively. Most likely we would choose fixed values for the parameters α, β, α^* and β^*.