1. **ALCOHOL DIGESTION**

(a) For female non-alcoholics the fitted regression line (or expected value of METABOL for a given GASTRIC) is obtained by setting FEM=1 and ALC=0. This yields Metabol = (-1.6597 + 1.4657) + (2.5142 - 1.6734)Gastric = -.1940 + .8408 gastric. For female alcoholics set FEM=1 and ALC=1 to find Metabol = (-1.6597 + 1.4657 + 2.5521 - 2.2517) + (2.5142 - 1.6734 -1.4587 + 1.1987) = .1064 + .5808 gastric.

(b) The full model (Model I) is the starting point. That model can be written as

\[ Y = \beta_0 + \beta_{\text{gascaric}} \text{GASTRIC} + \beta_{\text{fem}} \text{FEM} + \beta_{\text{alc}} \text{ALC} + \beta_{\text{fem.alc}} \text{FEM.ALC} + \beta_{\text{gast.fem}} \text{GAST.FEM} + \beta_{\text{gast.alc}} \text{GAST.ALC} + \beta_{\text{gast.fem.alc}} \text{GAST.FEM.ALC} + \epsilon. \]

Then we would like to test

\[ H_0: \beta_{\text{alc}} = \beta_{\text{fem.alc}} = \beta_{\text{gast.alc}} = \beta_{\text{gast.fem.alc}} = 0 \text{ vs } H_a: \text{one or more of these } \beta \text{'s is non-zero.} \]

Test statistic is

\[ F = \frac{(40.31 - 37.75)/4}{(1.57)} = .408. \]

We compare this to \( F_{4,24} \) distribution and find p-value > .50. No evidence against \( H_0 \) so it makes sense to act as if alcoholism status can be ignored (i.e., we can drop those terms). Note there was a typo here as the SSE for the reduced model should have been 40.81. It must be the case that the difference in SSE between the two models is equal to the difference in the SS(Model) between the two models.

(c) Wanted you to look at leverage (should be compared to \( 2*4/32 = .25 \) or \( 3*4/32=.375 \)), studentized residuals (values near 2 or higher are noteworthy but remember that we expect 5% or so in that range) and Cook’s distance (compared to \( F_{4,28,.5} = .863 \) or more generally any values near 1 are noteworthy). For the studentized residuals it is important to note that we expect some large values so anything above 2 does not automatically become an outlier. For full credit I wanted you also to note something (almost anything) about the specific observations that you were discussing. Observation 17 has high leverage (so potentially influential) but has a small residual (consistent with pattern of other observations) and thus not influential (low Cook’s distance). It has the highest gastric value for females. Observation 26 has a large negative residual but fairly ordinary values for leverage and Cook’s distance. There does not seem to be anything special about this case. Observations 31 and 32 both have high leverage, high residuals (especially obs 32) and high influence. Note that these are the two highest gastric levels and really stand out on the data plot and the residual plot.

(d) Models III and IV

i. This was worded somewhat awkwardly. All that I was aiming for here was an argument about why we might delete the two observations in this specific case. In this case we might justify deleting the two observations by noting that they are far from the other values of GASTRIC. Thus we could decide that we are only interested in studying the subpopulation with GASTRIC less than or equal to 3.0.

ii. Recall that each t-test is carried out assuming the other variables are included in the model. Thus FEMALE is not needed if you include GAST.FEM and GAST.FEM is not needed if FEMALE is included. This can happen when two variables are highly correlated as we might expect these to be.

(e) Models V and VI

i. A number of you did not sketch or draw the relationships implied by the two models. Model V has two parallel lines with the male line 1.53 units above the female line. Model VI has two lines that share an intercept (-.10) but have different slopes (male slope is 1.65 and female slope is 0.78).

ii. Neither model is a submodel of the other so we can’t use our usual testing approach. It is possible to compare using a model selection approach but not a formal test.
2. PACKAGE DELIVERY

(a) Model

i. Factorial means that each level of one factor is paired with each level of the other factors in the experiment.

ii. Let \( Y_{ijk} \) = delivery time for package \( k \) delivered by firm \( i \) at time \( j \). Then we usually write \( Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \) where \( \mu \) is the overall mean, \( \alpha_i \) is the firm \( i \) effect (with \( \sum_i \alpha_i = 0 \)), \( \beta_j \) is the time \( j \) effect (with \( \sum_j \beta_j = 0 \)), \( (\alpha\beta)_{ij} \) is the interaction effect (with \( \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0 \)), and \( \epsilon_{ijk} \) is the unique variation for package \( k \) (with \( \epsilon_{ijk}'s \) independent \( N(0, \sigma^2) \)). One could also have written this out as a regression model with indicator variables.

(b) Filling in the ANOVA table: DF: 2 for time, 2 for firm, 4 for firm*time, 45 for error, 53 total. Only one

(c) The time*firm interaction measures the degree to which the time effect is not the same for all firms

or equivalently that the firm effect is not the same for all times. For example, one firm may have an

(d) Contrast

i. Most everyone correctly specified weights as \( c = (-1, 2, -1) \) or something like that for testing the null hypothesis that \( -\mu_{morn} + 2\mu_{aft} - \mu_{eve} = 0 \) or equivalently that \( 2\mu_{aft} = \mu_{morn} + \mu_{eve} \) (which would mean the theory is wrong) against the alternative that \( 2\mu_{aft} > \mu_{morn} + \mu_{eve} \) which is consistent with the theory. When written in this way it assumes you will average the results for all firms together. It was also OK (and in fact a good idea) to express the weights in terms of the 9 treatment combinations, (-1, -1, -1, 2, 2, 2, -1, -1, -1). One thing to note is that if you want to interpret the final answer to be in “minutes” then you need to have the positive weights and negative weights sum to one, i.e., to use (-1/6, -1/6, -1/6, 1/3, 1/3, 1/3, -1/6, -1/6, -1/6). This ensures that you are contrasting two weighted averages of delivery times.

ii. The hard part in carrying out the contrast calculation is to make sure that you match the sample sizes that you use to define the contrast with the sample sizes that you use to calculate the standard error of the contrast. Thus if you work with \( \mu_{morn}, \mu_{aft}, \mu_{eve} \) then these are averages over 18 observations. If you work with a contrast defined on the cell means (i.e., \( \mu_{morn,A}, \mu_{morn,B}, \ldots \)) then these are averages over 6 observations. I find it easiest to work with the 9 treatment combinations and to make it easy to interpret would try to make the positive and negative weights sum to \( c \) and to make it easy to interpret would try to make the positive and negative weights sum to \( c \) and not be interpreted as “minutes” so one would say that you were 90% confident that the contrast was in this interval but couldn’t refer to it as the difference in delivery times.

(e) Regression

i. One advantage of using time as a continuous variable is that you can predict for any time, not just those in the study.

ii. One disadvantage of using time as a continuous variable is that it requires we model the relationship over 24 hours (linear, quadratic, etc.) when the coarser measure (morning, afternoon, evening) may be sufficient.
3. COST OF BUILDING NUCLEAR PLANTS

(a) Model 1

i. The variable cost can be skewed without violating the assumption. The distribution of cost that we see is the marginal distribution that is averaged over all the values of the covariates. The assumption is about the distribution for a fixed set of covariate values.

ii. The expected cost of a plant increases by 0.4 million for a 1 megawatt increase in capacity with all other variables held fixed.

iii. The probability of observing a test statistic as large or larger than the one we have observed (t = -1.56) if the true population coefficient is zero would be .12.

iv. If these were all correlated highly it could make interpreting the contributions challenging. We could assess whether this is happening by computing the variance inflation factor or closely examining the correlation matrix (though the VIFs are better). The intended focus was on diagnosing multicollinearity and thus talk of testing models is not really appropriate.

(b) Model 2

i. We are 95% confident that the mean log(cost) of the subpopulation of plants resembling the first plant is between 5.54 and 5.89. Exponentiating tells us that we are 95% confident that the median (?) cost for this subpopulation is between 255 and 361 million.

ii. (A) There are many negative residuals on both the left and right sides of the plot with many positive residuals in the middle. This suggests that there is a nonlinear relationship. (B) We could address this issue by including quadratic terms for the various (continuous) predictors.

iii. It turns out that these two approaches both produce the same estimated regression coefficients for the other variables. There is one extra parameter now to explain the one observation. It can be used to ensure that the residual there is zero and then the remaining parameters will produce the same minimization problem. This can be proved mathematically but I wasn’t expecting that. Intuition was sufficient. Here’s one somewhat formal approach. Let’s assume that observation n is the outlier and let I be an indicator for the outlier (I = 1 for case n and I = 0 for other cases). We can write the sum of squared errors for this model as the sum of two terms, one for the first n – 1 cases and one for the last case.

\[
\sum_{i=1}^{n}(Y_i - (X\beta)_i - \alpha I)^2 = \sum_{i=1}^{n-1}(Y_i - (X\beta)_i)^2 + (Y_n - (X\beta)_n - \alpha)^2.
\]

Note that the indicator is zero for the first term on the right so I have not included \( \alpha \) there. The indicator is one for the second term on the right. Now if we choose the remaining coefficients (\( \beta \)) to minimize the first term (this is equivalent to minimizing the SS with the case deleted), then we can choose the coefficient of the indicator (\( \alpha \)) to make this last term zero.
FUEL EFFICIENCY DATA ANALYSIS

• The analyses here were generally quite good. It is a bit of a tricky data set for reasons described below.

• One very important point is that given our interest in the effect of E10 gas on fuel consumption it is not really appropriate to just declare it not significant and stop there. As we have discussed many times the null hypothesis testing approach is appropriate when there is reason to prefer 0 effect as the explanation until an alternative is proven. That’s not the case here. We know that ethanol should require more fuel consumption. Whether the observed difference is statistically significant is largely a function of sample size. A good summary or conclusion should give the best estimate you can (and a confidence interval). The average value of the outcome measure is about 5 liters per 100km. The effect of ethanol is to increase this fuel usage by about 1% (or a little more); it costs about 5% less so maybe a good deal.

• One of my hints was not very helpful. I suggested that outside temperature and the difference between inside and outside temperature would matter. It turns out these two are very highly correlated because inside temperature is nearly constant. Thus you really only needed one of them.

• Almost everyone noticed that the basic regression using fuel consumption as the response led to a residual plot with a number of problems (non-constant variance, non-linearity and non-normality). There are several approaches to this:

  – The log transformation helped but didn’t really fix the problems. It is OK to go in this direction but you should address the fact that there is still non-constant variance, perhaps by using the sandwich estimator to get the robust standard error.

  – It turns out that using the transformation 1/y works much better. We don’t often think of this transformation but it makes perfect sense here. It then becomes a measure of distance per liter (which is how gas usage is measured in the U.S. and some other countries). Thus 5 liters per 100km becomes .2 (100km/liter) or 20km/liter.

  – A third option is to leave the response untransformed but use weighted least squares to estimate. The response is highly variable for short trips. This makes sense because the response is an average over distance and a short trip would be expected to have more variability. There are driving issues as well (cold car, cost of starting up is high, etc.). Using weighted regression with weight equal to distance (or speed since distance and speed were correlated) seemed to help solve the heterogeneity of variance problem.

• Regardless of which approach you took, it is also important to followup on the idea that the relationship of fuel consumption with speed was expected to be non-linear. People found lots of ways to explore this. Adding a quadratic term (and perhaps a cubic) is one crude approach that seems to help.

• One interesting question is whether it is OK to use distance as a predictor. Some people left it out because distance is part of the response measure. I think it is OK to include as a predictor (and if you do it seems like a non-linear relationship is best).

• This is a large data set so had some large outliers and some high leverage points. Nothing seemed particularly influential and when examining the data none of the values seemed particularly odd. Thus I do not believe removing observations was justified.

• One interesting approach taken by a few people was to separate the drives into two subpopulations based on distance. This complicates reporting the E10 effect since it can be different in the two subpopulations.