

1. HOSPITAL-ACQUIRED INFECTION DATA

- (a) Let μ_{med} denote the mean infection rate in the population of hospitals with medical schools and μ_{no} denote the mean infection rate in the population of hospitals without medical schools. We want to test $H_o : \mu_{med} = \mu_{no}$ vs $H_a : \mu_{med} \neq \mu_{no}$. The pooled t test statistic is $t = (5.09 - 4.22) / \sqrt{s_p^2(1/96 + 1/17)} = 0.87 / \sqrt{1.7176 * (1/96 + 1/17)} = 2.52$ where $s_p^2 = (95 * 1.34^2 + 16 * 1.12^2) / 111$. The p-value (two-sided) using the t_{111} distribution is $.01 < p < .015$ (rounding down to 60 degrees of freedom in the tables); since the p-value is so small we reject H_o . Note that hospitals with medical schools have the higher infection rate! It is OK to use the alternative (unpooled) procedure but then you have to be careful with the degrees of freedom; the conservative procedure is to use 16 (smaller sample size minus one) degrees of freedom.
- (b) Model 1
- The t statistic for “med” is -2.06 . It is best to round down to 60 degrees of freedom (this is the conservative thing to do) which yields $.04 < p < .05$ (two-sided).
 - A one-day increase in average stay is associated with a .2424 increase in expected infection rate for a hospital with all other predictors held fixed.
 - The “Root MSE” is an estimate of σ , the standard deviation of the regression errors (the ϵ 's). The “Std Error Residual” is an estimate of the standard deviation or standard error of the first residual ($e_1 = y_1 - \hat{y}_1$). They are related as “Std Error Residual” = “Root MSE” $\sqrt{1 - h_{11}}$.
 - It is possible to have Type II SS greater than Type I SS; it happens in Model 1 for a couple of variables. Most people just gave definitions here as explanation. That was very good but I was looking for a bit more. For this too happen there has to be a variable later in the sequence that makes the variable be more important.
- (c) The estimated effect of having a medical school is -0.66 with all other factors held fixed which means hospitals with medical schools have lower infection rates. Recall that in (a) hospitals with medical schools had higher infection rates! This is a large change. Obviously controlling for other factors has a big impact. Was a bit disappointed here with the focus on statistical significance – that’s missing the big picture.
- (d) Diagnostics
- The key point here is that we expect 5% of observations to have externally studentized residuals bigger than 2 in absolute value by chance and about 1/300 of observations to have externally studentized residuals bigger than 3 in absolute value by chance. Thus 6 observations out of 113 bigger than 2 is not unusual; the one observation with residual bigger than 3 is noteworthy but probably not unusual enough to require action. Remember don’t delete observations unless you have a very good reason.
 - The primary measure of influence is Cook’s distance. Leverage measures potential influence not actual influence. If a point has small Cook’s distance, then the regression coefficients will not change much if that observation is removed.
- (e) Effect of “cult”
- The effect of a 1 unit increase in “cult” depends on the region and the presence/absence of a medical school because of the interactions. For a hospital in the W with no medical school (the reference group) we just look at the coefficient of “cult” to see the expected change in infection rate is .00321. For a hospital in the W with medical school the slope is $.00321 - .02951 = -.02630$. In the S with no medical school the slope is $.00321 + .08447 = .08768$; in the S with medical school the slope is $.00321 + .08447 - .02951 = .05817$. Results for other regions are obtained in the same way.
 - Test $H_o : \beta_{med,cult} = \beta_{s,cult} = \beta_{ne,cult} = \beta_{nc,cult} = 0$ vs H_a that at least one is not-zero. Compare the full model (Model 3) vs the reduced model (Model 2) using $F = ((86.349 - 79.858)/4) / (79.858/100) = 2.032$ and p-value just above .10 (rounding down to $F_{4,60}$). There is thus only weak evidence against H_o and we would therefore probably choose not to reject H_o and rely on the simpler model without interactions (Model 2).
- (f) There are several factors that lead one to say the patient should not turn down an x-ray based on this analysis. Most important for me is that this is an analysis of hospital data; it is not about individual patients. Several people also observed the difficulty in drawing causal inference which was good but did not receive full credit. Talking about collinearity got some credit but this was not the main point.

2. MARKET SHARE

(a) Contrasts

- i. Define the contrast $\gamma = -0.5\mu_{std,std} + 0.5\mu_{dsc,std} - 0.5\mu_{std,enh} + 0.5\mu_{dsc,enh}$. Then our estimate is $\hat{\gamma} = -0.5 * 2.40 + 0.5 * 2.42 - 0.5 * 2.74 + 0.5 * 2.90 = 0.09$ and the estimated standard error of the contrast is $s.e.(\hat{\gamma}) = \sqrt{MSE(\sum_i c_i^2/n_i)} = \sqrt{.024656 * .117445} = .0538$. Note that MSE is obtained by pooling variances from the four groups $((8 - 1) * .12^2 + (7 - 1) * .11^2 + (8 - 1) * .18^2 + (13 - 1) * .18^2)/32$. The 95% CI is $.09 \pm t_{32,.975} * .0538 = .09 \pm 2.042 * .0538 = (-.02, .20)$ (rounding down to 30 d.f.).
- ii. We are 95% confident that the population mean market share under discounting exceeds the population mean market share under standard pricing by an amount between -.02 and .20. You don't interpret a CI by indicating whether it is significant or not; if we wanted that information we would do a statistical test.
- iii. This is subtle ... the language used here is describing an interaction. The "effect" of a price discount under standard advertising is $\mu_{dsc,std} - \mu_{std,std}$ and the effect of a price discount under enhanced advertising is $\mu_{dsc,enh} - \mu_{std,enh}$. If we want to see if these are different we should look at the difference between the two quantities, $\gamma = \mu_{dsc,enh} - \mu_{std,enh} - \mu_{dsc,std} + \mu_{std,std}$. This leads to consideration of a contrast with weights (1, -1, -1, 1).

(b) Economic growth

- i. My intention was that folks would think of residuals to check the model. Here one might plot residuals ($Y_{ij} - \bar{Y}_i$ for strategy i) versus economic growth G . Each month would have a residual and a value of growth. We would examine a scatterplot to look for a linear or other pattern. Many people mentioned that they would include growth in a model and test the coefficient - this would work too.
- ii. The best idea is to introduce indicators for the strategies (say D for discount, A for advertising, $D * A$ for their interaction) and use these indicators along with the variable growth in a regression model. Using subscript i to denote month i we can write $Y_i = \beta_0 + \beta_1 D_i + \beta_2 A_i + \beta_3 D_i A_i + \beta_4 G_i + \epsilon_i$. We could also include interactions. Don't forget to include the coefficients when you write down the model. Some people talked about two factor ANOVA, I guess assuming G_i was categorical. That could work. Others talked about blocking - this is possible but tricky because you are blocking after data was collected. Thus one would have to be careful in designing blocks.

3. ENERGY USE IN OFFICE BUILDINGS

(a) Model 1

- i. You need to address both plots. The first plot shows some evidence of increasing variance though several of you noted the small sample size on the left-hand side. The second plot looks pretty normal except for one positive outlier.
- ii. I like this question. Most people's instincts matched my own inclination which would be to keep the variable because it is telling us something even though it is counter to expectations. You needed to provide some reason for keeping it. Most likely it is correlated with some variable that we did not include and thus is a proxy for something else. Some people pointed out that we should check the data to see if one or two cases might be causing this result. An excellent idea that I had not thought of.

(b) Model 2

- i. Because the logarithmic transformation changes the scale of measurement it is not reasonable to compare MSEs; the MSE will generally be smaller for data after a logarithmic transformation (unless all the values of Y are really small!).
- ii. I was looking for the formal meaning of p -value here. The p -value is the probability of observing a regression coefficient as or more extreme than the observed coefficient (-.0001) if the true model has regression coefficient zero.

- (c) Discussions here were generally good but I was hoping for some kind of statement about the joint effect of these related variables. One should definitely start by noting that $E(\log Y) = 13.65 + .53 \log SA - .50 \log Heat - .06 \log InsulProp + .009AC - 1.046 \log DD - .37 \text{Stm}$. You can exponentiate and then claim $E(Y) \approx cSA^{.53}Heat^{-.50}InsulProp^{-.06}DD^{-1.05}e^{.009AC-.37\text{Stm}}$ but this is not quite right because you can't exponentiate and get back the mean. More precise to talk about the median. Any discussion like this got credit.