Solutions to HW7, Stat 401 Spring 2011

(1). Ch. 10 #18. Here we have $I = 5$ groups, consisting of $J = 4$ observations each, with respective group means and sample variances

$$\{\bar{X}_{i\cdot}\}_{i=1}^{5} \;=\; (12.75, 17.75, 17.50, 11.50, 10.00)$$

$$\{S_i^2\}_{i=1}^{5} \;=\; (17.583, 12.917, 4.333, 21.667, 15.333)$$

(a) Then $SSB = 200.3$, $SSW = 215.5$, $F = 3.485$, which has p-value .033 (and in any case the F-statistic is $> F_{4,15,.05} = 3.056$ and therefore significant.

(b) Putting the pairwise group comparisons $(\bar{X}_{i\cdot} - \bar{X}_{k\cdot})\sqrt{J/MSE}$ into a matrix gives

$$\begin{pmatrix} 0.000 & -2.638 & -2.506 & 0.660 & 1.451 \\ 2.638 & 0.000 & 0.132 & 3.298 & 4.089 \\ 2.506 & -0.132 & 0.000 & 3.166 & 3.957 \\ -0.660 & -3.298 & -3.166 & 0.000 & 0.791 \\ -1.451 & -4.089 & -3.957 & -0.791 & 0.000 \end{pmatrix}$$

In this setting the upper .95 quantile of the Tukey(5,15) studentized range statistic is 4.367, so only the pairwise comparisons with larger absolute-value studentized difference are interpreted as significantly different (although a couple are close, i.e. the group 2 vs. 5 and 3 vs. 5 comparisons). In other words, **none** of the groups are judged significantly different, even though we saw that the F statistic rejected the hypothesis that the means were all the same !

(2). Ch. 10 #20. We consider a setting with $I = 3$, $J = 5$, and $\{X_{i\cdot}\}_{i=3} = (10, 15, 20)$, so $SSTr = 5 \cdot (5^2 + 5^2) = 250$. The SSE's for which the ANOVA F test would reject are those for which $(250/2)/(SSE/12) > F_{2,12,.05} = 3.885$, i.e. for which $SSE < 386.07$. On the other hand, the maximum absolute difference for means beyond which the Tukey method finds at least two means different is $3.773\sqrt{SSE/(12 * 5)}$. With the means as given, the Tukey method finds groups differences if and only if $10 > 3.773\sqrt{SSE/(12 * 5)}$, or $SSE < 421.481$. So if SSE lies between 386.07 and 421.481, then ANOVA accepts but Tukey rejects the null hypothesis.

(3). Ch. 12 #6. The scatterplot shows a nearly linearly decreasing trend for the bulk of observations, those with $x < 150$. However, there are two observatins with large x values (roughly 178 and 188), which have very different corresponding y-values, and if only these two observations were seen, then we would have concluded that y was sharply *increasing* with x. If only one of these observations were in the dataset, it would definitely distort the estimated slope; but since both are in the dataset, the fitted slope will be about the same as if they were absent.

(4). Ch. 12 #12. (a). Here:

$$S_{xx} = 390995 - 517^2/14 = 371903, \quad S_{xy} = 25825 - 517 * 346/14 = 13048$$

$$S_{yy} = 17454 - 346^2/14 = 8903$$

Therefore $\hat{b} = S_{xy}/S_{xx} = 0.0351$ and $\hat{a} = 346/14 - (517/14) * .0351 = 23.418$. So the fitted equation is: $y = 23.418 + .0351 * x$.

(b). The prediction at $x = 35$ is 24.646, with residual (obtained using $y = 21$ from Ex. 12.4) equal to -3.646.

(c). $SSE = S_{yy} - S_{xy}^2/S_{xx} = 8445.2$, so $MSE = 8445.2/12 = 703.8$, and $\hat{\sigma} = \sqrt{703.8} = 26.53$.

(d). Proportion of variation explained by regression is $1 - SSE/SST = 1 - 8445.2/8903 = 0.05$.

(e). We are told to delete the last two observations, (103,75) and (142,90). The new summary quantities are:

$$S_{xx} = 2156.7, \quad S_{xy} = 1217.3, \quad S_{yy} = 998.9, \quad \bar{x} = 22.67, \quad \bar{y} = 15.08$$

The new fitted regression coefficients are $\hat{b} = 0.564$, $\hat{a} = 2.290$, and the new $\hat{r}^2 = 1217.3^2/(998.9 * 2156.7) = .688$. Thus the deletion of the last two observations makes a huge difference to the model fit !

(5). Ch. 12 #20. We use the MINITAB output to avoid doing new calculations wherever possible. (a). The least squares estimates are the estimated coefficients, $\hat{a} = .3651$, $\hat{b} = .9668$. (b). The prediction is $.3651 + .9668 * .5 = .8485$. (c). $\hat{\sigma} = (MSE)^{1/2} = .193$. (d). $SST = 1.4533$, of which a percentage $\hat{r}^2 = SS.Regr/SST = .717$ is explained by regression.

(6). Ch. 12 #34. (a). *Model utility test* is the name for the t-test of $b = 0$, equivalent to the F-test comparing $t_{11}^2 = F_{1,11} = 27.94$ to $F_{1,11,.05} = 4.84$. The test rejects, with very small p-value $< .0003$.

(7). Ch. 12 #36. (a) The scatterplot looks reasonably linear for the first 6 observations, but the great distance of the last point from the others makes one doubt the appropriateness of analyzing all 7 data points together. (b) Regression analysis gives $\hat{b} = .000621$, $\hat{r} = 0.9647$, so the proportion of variation attributed to regression is $\hat{r}^2 = .9307$.

(b). The wording makes it seem that the author wants us to test the hypothesis $H_0 : \; b * 900 \geq .6$ versus $H_A : \; b * 900 < .6$. The auxiliary calculations are: $S_{xx} = 508479.4$, $\hat{\sigma^2} = .00292$. Then the t test statistic for this test rejects for values $t_5 < -2.015$, but the calculated statistic is $(.000621 - .000667) * (508479/.00292)^{1/2} = -.607$. So **no** the evidence is not substantial that the average increase is $< .6$.

(c). We are asked for a confidence interval about b; the 95% interval is $= .000621 \pm 2.571 * (.00292/508479)^{0.5} = (.00601, .00640)$.