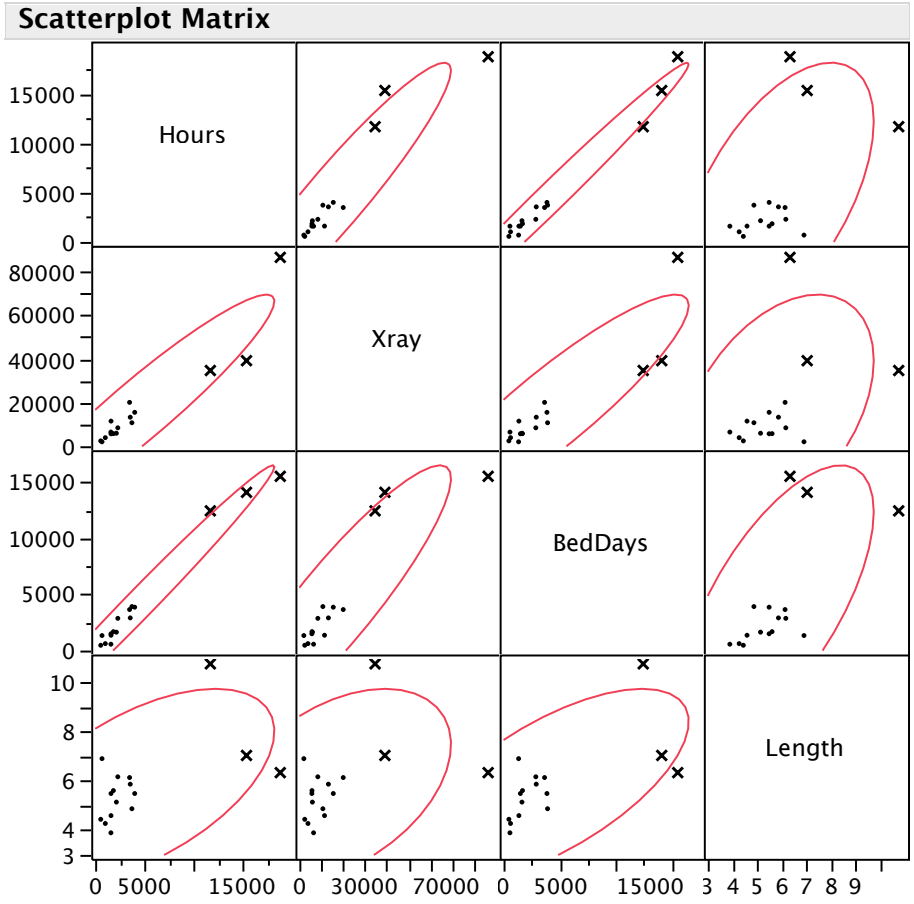
Insurance 260
Spring, 2009

Solutions, Assignment 2

4.2 The following shows the scatterplot matrix of the hospital data. Hours should be in the first row to highlight the response variable. The ellipses convey the correlations. We can see several leveraged outliers (large hospitals) in these plots. These are high-lighted in the graph as "x"s and are the last 3 cases in the data table. The linear association between Hours and X-ray and Hours and bed days appears fine, albeit for the leverage points. Length is less clear.

The intercept β_0 controls the level of the fit and measures hours needed when all 3 explanatory variables are zero (a bit of an extrapolation). We'd expect it to be near zero, though positive since the building would require, for example, maintenance regardless of patients. Individual slopes measure the partial effects of each explanatory variable. For example β_1 for x-rays measure the average increased hours of labor per additional x-ray among hospitals with given counts of bed days and length of patient stays.



4.4

(a) This output summarizes the multiple regression. The text does not ask for many plots, but I've shown two useful plots. The 3 leveraged cases are again very apparent. The estimated intercept sets a baseline of about 1947 hours of labor on average to maintain an empty hospital without patients. Each additional x-ray exposure adds about 0.039 of an hour (about 2.5 minutes) of labor. (Seems small to me.) Another bed day adds on average 1.04 hours of labor. Longer average stays imply on average less labor for a given number of x-rays and bed-days; evidently patients with longer stays do not require as many hours of daily care. Notice that the partial slope for length is negative even though the marginal association is positive; this happens due to correlation among the explanatory variables (collinearity). Note the very large values of VIF for X-rays and BedDays; these two are also highly correlated.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	1946.802	504.1819	3.86	0.0023*	.
Xray	0.0385771	0.013042	2.96	0.0120*	7.8283193
BedDays	1.039392	0.067556	15.39	<.0001*	11.396195
Length	-413.7578	98.59828	-4.20	0.0012*	2.5195594

(b) To get the prediction, just plug the indicated values into the estimated regression:
1946.8 + 0.03858*(56194) + 1.0394*(14077.88) - 413.76*(6.89) ≈ 15896 hours

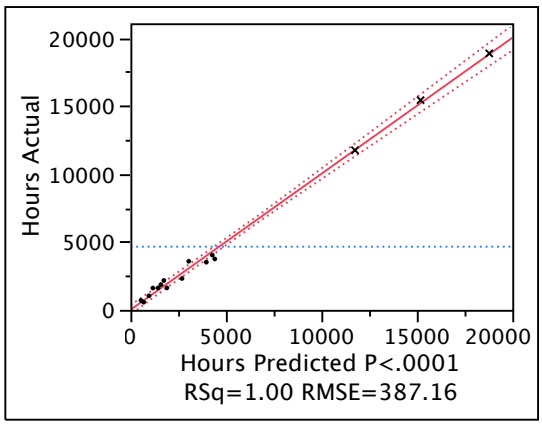
(c) skip

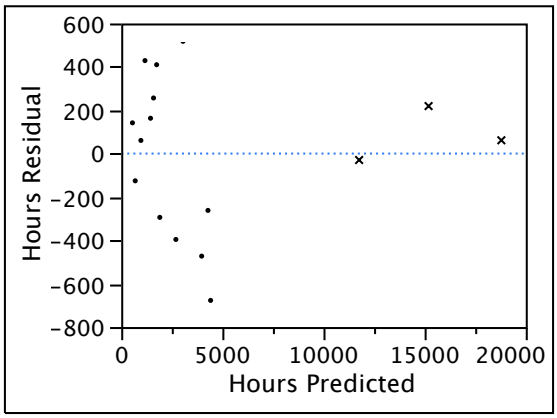
Summary of Fit

RSquare	0.996125
RSquare Adj	0.995156
Root Mean Square Error	387.1598
Mean of Response	4643.147
Observations (or Sum Wgts)	16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	462327889	154109296	1028.131
Error	12	1798712	149892.68	Prob > F
C. Total	15	464126602		<.0001*





4.6

(a) skip. In case you're interested... s (the RMSE), s^2 and SSE are in the summary of the fit found in the Anova table shown on the previous page. s^2 (square of the RMSE) is in the table under the name "Mean square, error". Its calculated as

SSE / df(residual or error) = 1798712/(16-1-3) ≈ 150,000 hours

(b) skip

(c) R^2 = 0.996 and adjusted R^2 = 0.995. The easier way to compute the adjusted R^2 is to work with the unexplained variation, computing (as in the lecture notes)
adjusted R^2 = 1 - (SSE/(n-1-k))/(SST/n-1) = 1-(1798712/12)/(464126602/15) ≈ 0.9951

(d) skip

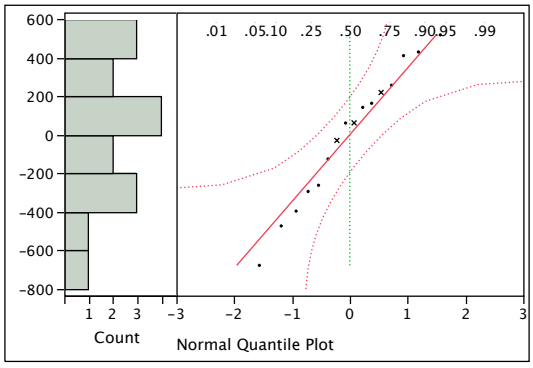
(e) skip

(f) skip

(g) The p-value for the F-statistic is in the Anova table. Since F = 1028, the p-value is far less than any reasonable α level; reject the null hypothesis H0: $\beta_1 = \beta_2 = \beta_3 = 0$. There's clearly linear association between this collection of explanatory variables and the number of labor hours.

4.8 (b) At the α = 0.05 level, all 3 explanatory variables are statistically significant. If we lower α = 0.01, then the number of X-rays is not.

4.10 The predicted number of hours is less than the actual by 17207.31-15896=1311.31 hours. The RMSE is about 387 hours, so this is exceptionally far from the predicted value, lying about 1311/387 ≈ 3.4 RMSEs above the prediction. That's way outside the 95% prediction interval which runs from about $\hat{y} \pm 2$ RMSE for example. (The 95% prediction interval from JMP for this case is from 14906 to 16886 hours. This hospital is using far more labor hours than would be anticipated from the regression. (It would be good to check for normality before doing this, however. With so little data, these qualify as "possibly normal", but a larger sample would be needed to check the assumption with some chance of rejecting normality.)



4.19 "Promotions run during day games and weekends have greater impact on attendance." To address this claim, first off notice that the presence of other variables, such as temperature and the nature of the opponent, remove these possible confounding variables from the analysis. (For example, games are likely to be warmer during the day comfortable fans might respond differently to promotions, but this model removes the effect of temperature from the comparison of day and night games.)

The interaction 5059 Promotion*DayGame implies that the combination of a promotion with a day game adds, on average, about 5,059 fans beyond what we would expect for a promotion or a day game considered separately. Consider the part of the model dealing with promotions and day/night games:

fit = baseline + 4745 Promotion - 424 DayGame + 5059 Promotion*DayGame

Promotions during a day game add 4745-424+5059=9,380 on average, whereas promotions during a night game add "only" 4,745.

For the weekend effect, the relevant part of the fit is

fit = baseline + 4745 Promotion +4845 Weekend - 4690 Promotion*Weekend

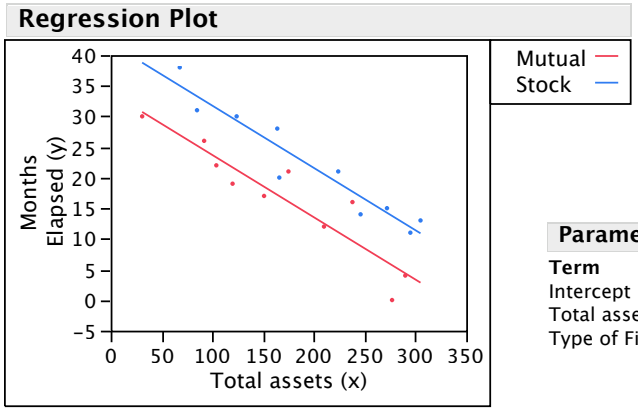
Promotions on the weekend, "holding fixed the other variables," add about 4745+4845-4690 = 4900 on average to attendance. Promotions on weekdays add 4,745. It does not appear that promotions add much during weekend games. (Weekend games are already high.)

Advice: Offer promotions to attract people to otherwise poorly attended games. No need to compete with yourself.

4.20

(a) The fits look vaguely parallel, so we may not need an interaction. The shift provided by adding the dummy variable to the regression may be all that is needed.

(b) As fit here, the dummy variable indicates mutual companies. Given two companies of equal assets (size), the model predicts the mutual company to adopt the innovation about 4 months more quickly. The estimate is -4.03 and is highly statistically significant.



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	37.901804	1.770041	21.41	<.0001*
Total assets (x)	-0.101742	0.008891	-11.44	<.0001*
Type of Firm[Mutual]	-4.027735	0.729553	-5.52	<.0001*

(c) We reject H_0 at any reasonable α level. The 95% confidence interval for the dummy variable is $b_2 \pm t_{0.025,17}\, SE(b_2)$ = -4.028 ± 2.11*0.7296 ≈ -5.57 to -2.49 months

(d) The interaction is not statistically significant; it appears that the effect of size on months to adoption is the same for both stock and mutual companies.

4.22 Here's the JMP version of the second-degree model with the dummy variables.

(a) Campaign C is beating the other two. Campaign A is coming in lower than A by 0.38 (hundred thousand) and Campaign B is lower by about half that at 0.17 (hundred thousand). Campaign A is lower than Campaign B by about 0.38 -0.17 = 0.21 (hundred thousand).

(b) JMP will do the prediction. Its formula gives $\hat{y}$ = 8.50 with 95% prediction interval from about 821,300 to 878,800.

(c) skip

The t-statistics are not the right approach here (though the outcome is similar in this example). To test whether both dummy variables add value **simultaneously**, use the effect test for Campaign. It is shown below and indicates that F = 19.6 with p-value quite small. We can easily reject the null hypothesis and conclude that the campaigns are significantly different in performance at any reasonable α level.

The reason for avoiding the t-statistics in this case is that in later examples (such as the dummy variable for month of the year), you will have 11 dummy variables. Some will be significant and others not. You'd like to get an overall assessment of the net differences among categories, not determine whether some are significant. Plus, we 11 dummy variables, what is the probability of getting a significant t-stat purely by chance: after all, you have eleven tests.

Indicator Function Parameterization

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	25.994472	4.786568	23.00	5.43	<.0001*
Advertising Expense (x3)	-6.537671	1.581367	23.00	-4.13	0.0004*
Price Diff (x4)	9.0586843	3.031705	23.00	2.99	0.0066*
Advertising Expense (x3)*Advertising Expense (x3)	0.5844439	0.129872	23.00	4.50	0.0002*
Advertising Expense (x3)*Price Diff (x4)	-1.156481	0.455736	23.00	-2.54	0.0184*
Campaign[a]	-0.381776	0.061253	23.00	-6.23	<.0001*
Campaign[b]	-0.16809	0.063707	23.00	-2.64	0.0147*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Advertising Expense (x3)	1	1	0.29246430	17.0915	0.0004*
Price Diff (x4)	1	1	0.15277359	8.9280	0.0066*
Advertising Expense (x3)*Advertising Expense (x3)	1	1	0.34653331	20.2513	0.0002*
Advertising Expense (x3)*Price Diff (x4)	1	1	0.11018966	6.4395	0.0184*
Campaign	2	2	0.67082865	19.6015	<.0001*