Problem 1

a) Randomization means that each patient is randomly allocated to treatment or placebo. If that isn't done, systemtaic differences between the two groups might interfere with the interpretation.

b) Placebo: (1783.4,2547.2) Treatment: (13.92.8,1925.6). The fairly large number of patients in the two groups means that the assumptions behind the intervals are approximately satisfied.

c) $\sqrt{(57 * 2165.3^2 + 54 * 1012.0^2)/111} = 1256.9$. Clearly a one-sided situation. $t = (1659.2 - 2165.3)/(2156.9\sqrt{1/55 + 1/58}\ ) = -2.13$ with P-value 1.7% with df=111.

Problem 2

a)

	sums of squares	df	mean squares	F-value
Treatments	338.8	2	169.4	6.6
Error	307.6	12	25.6	
Total	646.4			

and the F-value is signiicant at around 1%. It is wasteful to exclude the former sale which must lead to a less sensitive analysis, yet the significane, if found as here, is safe. Another weakness is that we would always be interested in estimating the numerical effect of the sales promotion techniques which means that regression must be a more natural approach.

c) The best sales promotion is the first one and the differences to the two others are convincingly signiicant from the printout. Predictions are 35.9, 31.8 and 25.2.

d) We have now corrected for the natural sales at the various shops. That was part of the random error earlier.

e) The point is that the regression coeffient for x is positive. This is a safe conclusion since the confidence interval excludes 0. It's about $(0.7, 1.1)$.

Problem 3

The expected numbers in each cell are computed as a product of the two corresponding row and column totals, divided by the grand total 4526. For example, the upper left number is given as $\dfrac{2691 \cdot 1755}{4526} = 1043.461$

The test statistic is calculated as follows:

$$\chi^2 = \frac{(1198 - 1043.461)^2}{1043.461} + \cdots + \frac{(1278 - 1123.461)^2}{1123.461} = 91.610$$

Under the null hypothesis is the test statistic χ^2 approximately chi-square distributed with 1 degree of freedom. Hence a 1% test would reject if $\chi^2 \geq 6.63$. This is a very clear rejection of the null hypothesis of equal admission probabilities for the two genders, where the data show a bias in favor of male applicants.

b) The applicants are either admitted or rejected, so the response is a binary variable. The probability of admittance is assumed to be a function of the factors `gender` and `prog`. Hence a logistic regression model can be used if we specify this probability on the logistic form. Further, we have to assume that the applicants are treated independently of each other.

If gender is the only covariate, then the model can be written:

$$p(x) = P(y = 1|x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}, \quad x = 0, 1$$

c) The odds ratio for male versus female applicants is

$$OR = \frac{\frac{p(1)}{1-p(1)}}{\frac{p(0)}{1-p(0)}} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$$

The estimate is hence $\exp(\hat{\beta}_1) = \exp(0.61035) = 1.841$.

From the 2×2 table we estimate the probabilities for admitting, respectively, males and females to be

$$1198/2691 = 0.4452, \quad 557/1835 = 0.3035$$

Thus we obtain the odds ratio

$$OR = \frac{\frac{0.4452}{1-0.4452}}{\frac{0.3035}{1-0.3035}} = 1.841$$

which is the same as we obtained via the logistic regression. (This is because in a model where program is not part of the data, the logistic model is saturated. See also last subproblem of the current Problem).

d) The present model has `gender` and `prog` as covariates (factors). It has apparently much better fit than the previous one due to a much lower deviance. In the new model, gender is no longer significant, which indicates that the difference in admittance probability is dominated by differences between programs.

We now want to test formally the null hypothesis that the 5 β-coefficients corresponding to `prog` are all 0. The test statistic is

$$G = D_0 - D$$

where D_0 is the deviance under the null hypothesis and D is the deviance in the model which includes the programs. Thus, from the R-outputs, $G = 783.61 - 20.20 = 763.41$. Under the null hypothesis we would have that G is approximately chi-square distributed with $df = 5$. The conclusion is therefore a very clear rejection, giving a conclusion that the factor `prog` has a significant influence on the probabilities of admittance.

e) We first test the null hypothesis that there is no interaction between the factors `gender` and `prog`. As in the previous subproblem, this is done by considering the difference between deviances. Here, with the obvious meaning in the present case,

$$G = D_0 - D = 20.204 - 0 = 20.204$$

Under the null hypothesis is G approximately chi-square distributed with $df = 5$, due to 5 less parameters under the null hypothesis. The p-value is hence $P(\chi^2_5 > 20.204) = 0.0011$, so we conclude a significant interaction. (The attached tables will show that we reject at any reasonable significance level).

The main effect of `gender` is now estimated by the β-coefficient -1.0521, which shows a bias *towards* admitting women.

Let in the following x_1 be defined as in subpoint b), and let x_2 to x_6 be the indicators for the programs $B - F$, respectively, A being the reference program. Thus, e.g., $x_2 = 1$ if a student applied to program B and $x_2 = 0$ otherwise, etc.

Then the odds for a student with covariate vector $(x_1, \ldots, x_6)$ is the exponential of

$$\begin{aligned} \log \text{odds} &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 \\ &+ \beta_7 x_1 x_2 + \beta_8 x_1 x_3 + \beta_9 x_1 x_4 + \beta_{10} x_1 x_5 + \beta_{11} x_1 x_6 \end{aligned}$$

For a male applicant ($x_1 = 1$) this becomes

$$\begin{aligned} \log \text{odds} &= \beta_0 + \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 \\ &+ \beta_7 x_2 + \beta_8 x_3 + \beta_9 x_4 + \beta_{10} x_5 + \beta_{11} x_6 \end{aligned}$$

while for a female applicant ($x_1 = 0$) we get

$$\log \text{odds} = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

Thus the log odds-ratio for a male with respect to a female applicant is

$$\log \text{OR} = \beta_0 + \beta_7 x_2 + \beta_8 x_3 + \beta_9 x_4 + \beta_{10} x_5 + \beta_{11} x_6$$

so the odds ratio is

$$\text{OR} = \exp(\beta_0 + \beta_7 x_2 + \beta_8 x_3 + \beta_9 x_4 + \beta_{10} x_5 + \beta_{11} x_6)$$

The OR can now be computed (estimated) for each of the programs, as follows:

$$\begin{aligned}
\text{Program A:} &\quad \exp(-1.0521) = 0.349 \\
\text{Program B:} &\quad \exp(-1.0521 + 0.8321) = 0.803 \\
\text{Program C:} &\quad \exp(-1.0521 + 1.1770) = 1.133 \\
\text{Program D:} &\quad \exp(-1.0521 + 0.9701) = 0.921 \\
\text{Program E:} &\quad \exp(-1.0521 + 1.2523) = 1.221 \\
\text{Program F:} &\quad \exp(-1.0521 + 0.8632) = 0.828
\end{aligned}$$

Thus for program A the OR equals $\exp(\beta_1)$, estimated to 0.349. To obtain a 95% confidence interval we first find the standard 95% confidence interval for β_1 using the R-output:

$$-1.0521 \pm 1.96 \cdot 0.2671 = (-1.5756, -0.5286)$$

The 95% confidence interval for OR is obtained by exponentiating each side of this interval, to give

$$(0.2069, 0.5894)$$

f) The saturated model for the full data set used in the present Problem consists in representing each line in the data table (i.e., each possible combination of the factors) by a separate probability p_k for admission. For this we would need 12 parameters (probabilities). On the other hand, the model behind the last R-output, also has 12 parameters, $\beta_0, \beta_1, \ldots, \beta_{11}$. It can be seen that the $p_1, \ldots, p_{12}$ can be uniquely computed from the $\beta_0, \ldots, \beta_{11}$ (and vice versa). Thus the last model is equivalent to the saturated model, and hence has deviance equal to 0.

END