1. (40 points)

The file *hospital* contains data for personal health care expenditures by state for the year 1982. The variable *expadm* contains the average expense per admission into a community hospital, the variable *los* contains the average length of stay and the variable salary contains the average per employee in 1982. We are interested how the latter two variables affect the average expense per admission.

(a) Draw scatterplots of variables *expadm* vs. *los*, and of variables *expadm* vs. *salary*

(b) Using *expadm* as response variable and *los* as the explanatory variable, write down the least-squares regression equation.

(c) Suppose the average length of stay per admission at Fairyland Community Hospital is 6 days. Construct a 95% prediction interval for the average expense per admission at this hospital.

(d) Is there a significant linear relationship between expense per admission and the length of stay?

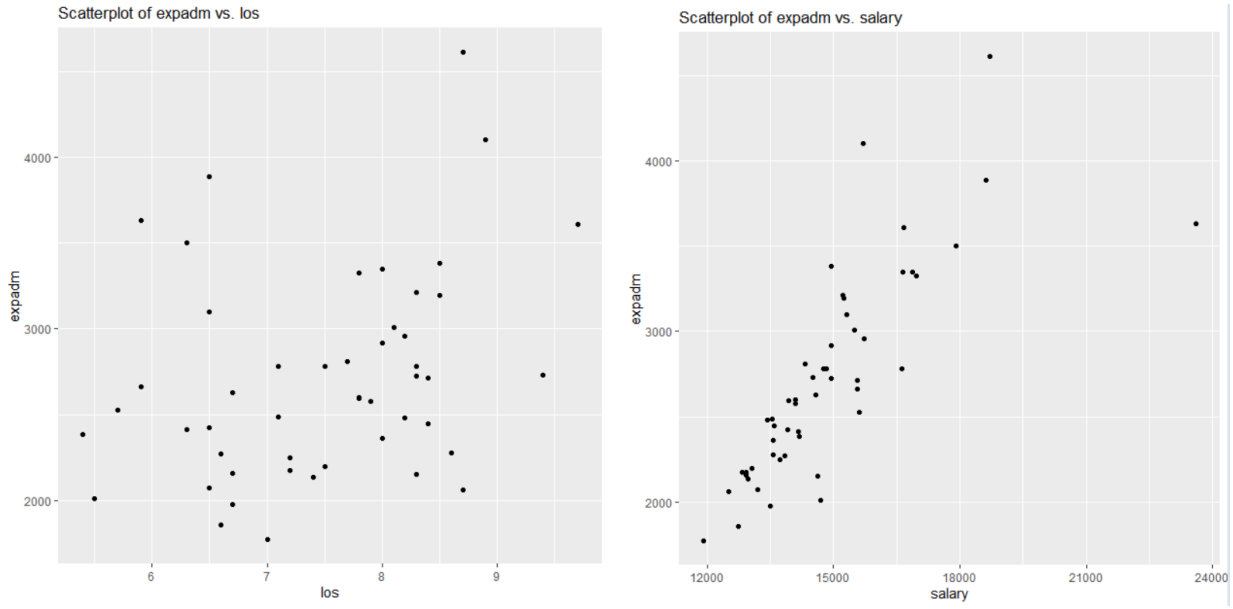
(e) Evaluate the fit of the model to data by creating a plot of residuals and a normal probability plot of the residuals. Discuss your findings.

(f) Fit the linear regression model where *expadm* is the response variable and *los* and *salary* are the explanatory variables. Interpret the estimated regression coefficients.

(g) What happens to the estimated coefficient of length of stay when average salary is added to the model?

(h) Does the inclusion of salary in addition to average length of stay improve your ability to predict mean expense per admission? Explain

Solution: (a)



(b)

```
> # --------------------(b)----------------------
> # Fit the linear regression model
> model <- lm(expadm ~ los, data = hospital)
> # Print the model summary
> summary(model)

Call:
lm(formula = expadm ~ los, data = hospital)

Residuals:
   Min     1Q Median     3Q    Max
-889.6 -428.1 -102.1  265.8 1663.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1281.96     608.10   2.108   0.0402 *
los           191.56      80.47   2.381   0.0212 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 577.6 on 49 degrees of freedom
Multiple R-squared:  0.1037,    Adjusted R-squared:  0.08538
F-statistic: 5.668 on 1 and 49 DF,  p-value: 0.02121
```

Using *expadm* as response variable and *los* as the explanatory variable, the least-squares regression equation is as follows:

$$expadm = 1281.96 + 191.56 * los$$

(c)

```
> # --------------------(c)----------------------
> # Predict the average expense per admission at Fairyland Community Hospital
> newdata <- data.frame(los = 6)
> pred <- predict(model, newdata, interval = "prediction", level = 0.95)
> pred
       fit      lwr      upr
1 2431.338 1234.791 3627.884
```

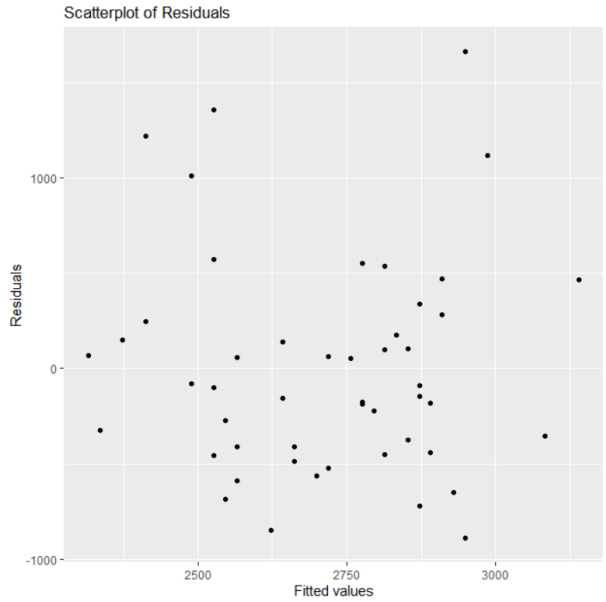
95% prediction interval for the average expense per admission at Fairyland Community Hospital is [1234.791 , 3627.884]

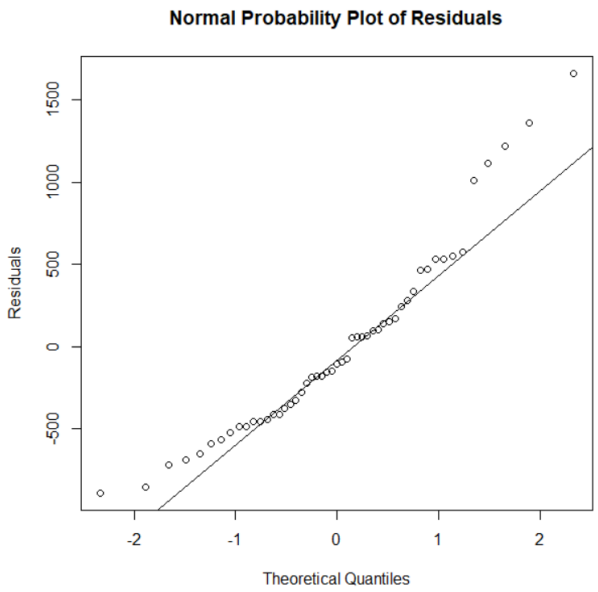
(d)

```
> # --------------------(d)----------------------
> # Test the null hypothesis that there is no linear relationship between expadm and los
> # Using a significance level of 0.05
> p_value <- summary(model)$coefficients[2, 4]
> if (p_value < 0.05) {
+   cat("There is a significant linear relationship between expense per admission and the length of stay.\n")
+ } else {
+   cat("There is no significant linear relationship between expense per admission and the length of stay.\n")
+ }
There is a significant linear relationship between expense per admission and the length of stay.
```

It can be concluded that, at 0.05 level of significance, the length of stay has a significant linear relationship with expense per admission. It length of stay increase by 1 unit, expense per admission with increase by 191.56.

(e)





The scatterplot of residuals vs fitted values reveals that the residuals have a random scatter around a horizontal line, but a few values are greater than 1000, indicating the presence of outliers and demonstrating that the linear model is not an adequate fit for the data. The absence of patterns in the residuals indicates that the assumptions of linearity, constant variance, and normality of errors have not been significantly violated.

The residuals' normal probability plot reveals that the residuals are substantially normally distributed since they follow a straight line. However, a few outliers stray from the normal distribution, suggesting that there may be some deviations from the mean. Overall, the plot indicates that the model is not a good match for the data, and the normality of errors may have space for improvement.

(f)

```
> # --------------------(f)----------------------
> # Fit the linear regression model
> model_lm <- lm(expadm ~ los + salary, data = hospital)
> summary(model_lm)

Call:
lm(formula = expadm ~ los + salary, data = hospital)

Residuals:
    Min      1Q  Median      3Q     Max
-920.43 -134.15    0.57  133.77  876.74

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2582.7364   464.7700  -5.557 1.18e-06 ***
los           213.7967    42.2077   5.065 6.45e-06 ***
salary          0.2490     0.0218  11.422 2.73e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 302.6 on 48 degrees of freedom
Multiple R-squared:  0.7589,    Adjusted R-squared:  0.7489
F-statistic: 75.55 on 2 and 48 DF,  p-value: 1.485e-15
```

The linear regression model with *expadm* as the response variable and *los* and *salary* as the explanatory variables can be written as:

$$expadm = -2582.7364 + 213.7967 * los + 0.2490 * salary$$

The estimated intercept is -2582.7364, which represents the expected value of *expadm* when both *los* and *salary* are equal to zero. The estimated coefficient for *los* is 213.7967, which indicates that for every unit increase in length of stay, we expect *expadm* to increase by 213.7967 units, holding *salary* constant. The estimated coefficient for *salary* is 0.2490, which indicates that for every unit increase in *salary*, we expect *expadm* to increase by 0.2490 units, holding *los* constant.

(g) When average salary is added to the model, the estimated coefficient of length of stay changes from 191.56 to 213.80. This indicates that the effect of length of stay on expense per admission increased after accounting for the average salary.

(h) Yes, the inclusion of salary in addition to average length of stay improves the ability to predict mean expense per admission. This is because the multiple R-squared value increased from 0.1037 in the model with only length of stay to 0.7589 in the model with both length of stay and salary. The adjusted R-squared value also increased from 0.08538 to 0.7489, which indicates that the added variable *salary* improved the goodness of fit of the model. Additionally, the *p-value* for both length of stay and salary are very small and significant, indicating that both variables have a significant effect on expense per admission.

2. (30 points)

Insurance adjusters from the Shrewd Insurance Company are concerned with the high estimates they are receiving from Garage Elegance for auto repairs compared to Joe's Garage. To verify their suspicions, each of the 15 cars recently involved in an accident was taken to both garages for separate estimates of repair costs. The data is shown below. (The repair estimates are in Hundreds of Dollars.) Also can find data in *Garage.txt*

CAR	GARAGE ELEGANCE	JOE'S GARAGE

1	**7.6**	**7.3**
2	**10.2**	**9.1**
3	**9.5**	**8.4**
4	**1.3**	**1.5**
5	**3.0**	**2.7**
6	**6.3**	**5.8**
7	**5.3**	**4.9**
8	**6.2**	**5.3**
9	**2.2**	**2.0**
10	**4.8**	**4.2**
11	**11.3**	**11.0**
12	**12.1**	**11.0**
13	**6.9**	**6.1**
14	**7.6**	**6.7**
15	**8.4**	**7.5**

Find the appropriate statistical method to solve this problem. Report your analysis. What can you say to the insurance adjusters about your result?

Solution: Based on the data provided, we can perform a two-sample t-test to compare the average repair cost estimates from Garage Elegance and Joe's Garage.

The null hypothesis is that there is no significant difference in the average estimates between the two garages.

The alternative hypothesis is that the average estimate from Garage Elegance is higher than that from Joe's Garage.

```
> t.test(garage$GarageElegance, garage$JoeGarage, alternative = "greater", paired = TRUE)

        Paired t-test

data:  garage$GarageElegance and garage$JoeGarage
t = 6.0234, df = 14, p-value = 1.563e-05
alternative hypothesis: true mean difference is greater than 0
95 percent confidence interval:
 0.4339886       Inf
sample estimates:
mean difference
      0.6133333
```

The *p-value* of the t-test is 0.0000156, which is less than the common significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there is enough evidence to suggest that the average repair cost estimate from Garage Elegance is significantly higher than that from Joe's Garage.

In other words, the statistical analysis provide strong evidence to support the insurance adjusters' concern about the high estimates from Garage Elegance compared to Joe's Garage. However, it's important to note that the sample size is relatively small, and additional data could provide more reliable results.

3. **(30 points) Groundhog Day.**
   (From https://www.ncdc.noaa.gov/customer-support/education-resources/groundhog-day)
   Every February 2, thousands gather at Gobbler's Knob in Punxsutawney, Pennsylvania, to await the spring forecast from a special groundhog. Known as Punxsutawney Phil, this groundhog will emerge from his simulated tree trunk home and look for his shadow, which will help him make

his much-anticipated forecast. According to legend, if Phil sees his shadow the United States is in store for six more weeks of winter weather. But, if Phil doesn't see his shadow, the country should expect warmer temperatures and the arrival of an early spring. But does Phil really know about the weather? The following table gives the data for the 20 years from 1988 to 2017. (There are more data on the web link. But for this problem, restrict the analysis to the following table only.)

	Phil sees its shadow	
March Temperature	**Yes**	**No**
Above normal	**16**	**8**
Below Normal	**5**	**1**

(a) State the appropriate null hypothesis and the alternative hypothesis.
(b) I would like to use the χ^2-test to solve this problem. However, this is a one-sided problem while the χ^2-test is usually used as a two-sided test. What should I do?
(c) Carry out the test at $\alpha = 0.05$ level.
(d) How can we use Groundhog Day's information to predict the March temperature of that year?

Solution: (a) Null Hypothesis: There is no association between Punxsutawney Phil seeing his shadow and March temperature being above or below normal.

Alternative Hypothesis: There is an association between Punxsutawney Phil seeing his shadow and March temperature being above or below normal.

(b) Since this is a one-sided problem, we can divide the chi-square test statistic by 2 to get the p-value for a one-sided test. Alternatively, we can use the binomial distribution to directly calculate the one-sided p-value.

(c)

```
> march_table
              Yes No
Above Normal   16  8
Below Normal    5  1
>
> # perform chi-squared test
> prop.test(march_table, correct = FALSE)

        2-sample test for equality of proportions without continuity correction

data:  march_table
X-squared = 0.63492, df = 1, p-value = 0.4256
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.5195005  0.1861672
sample estimates:
   prop 1    prop 2
0.6666667 0.8333333

Warning message:
In prop.test(march_table, correct = FALSE) :
  Chi-squared approximation may be incorrect
```

The p-value is 0.4256, and the test statistic is 0.63492 with 1 degree of freedom. We are unable to reject the null hypothesis since the p-value is greater than α. Thus, there is insufficient information to conclude that Phil's ability to forecast the March temperature is more accurate than chance.

(d) While the results of the chi-squared test do not provide evidence that Phil's predictions are better than chance, we can still use his predictions to make a rough prediction for the March temperature. If Phil sees his shadow, we can predict that the temperature will be below normal with a probability of 8/24 or 1/3. If Phil does not see his shadow, we can predict that the temperature will be above normal with a probability of 16/24 or 2/3. However, it is important to note that this is only a rough prediction, and other factors can influence the March temperature.

4. (30 points)

To demonstrate the effect of nematodes (microscopic worms) on plant growth, a botanist prepares 16 identical planting pots and then introduces different numbers of nematodes into the pots. A tomato seedling is transplanted into each plot. Here are data on the increase in height of the seedlings (in centimeters) 16 days after planting.

Nematodes	Seedling Growth			
0	10.8	9.1	13.5	9.2
1000	11.1	11.1	8.2	11.3
5000	5.4	4.6	7.4	5.0
10000	5.8	5.3	3.2	7.5

(a) I would like to know if the nematodes have any effect on the plant growth at all. Carry out appropriate statistical analysis. What is your conclusion?

(b) Before conducting the experiment, I decided to test whether the introduction of nematodes reduces plant growth. State an appropriate contrast to test this hypothesis and carry out the appropriate test.

(c) From the data, it seems that the biggest drop in plant growth is between the group treated with 1000 nematodes per plant and the group treated with 5000 nematodes per plant. If I then decide to test whether these two groups do have different plant growth, what is the appropriate contrast for this test? Carry out the analysis.

Solution: (a) To test whether the nematodes have any effect on plant growth, we can use a one-way ANOVA (analysis of variance) with the nematode treatment groups as the factor.

The null hypothesis is that the mean growth is the same across all nematode treatment groups.

The alternative hypothesis is that the mean growth is not same across all nematode treatment groups.

```
> group1 <- c(10.8, 9.1, 13.5, 9.2)
> group2 <- c(11.1, 11.1, 8.2, 11.3)
> group3 <- c(5.4, 4.6, 7.4, 5.0)
> group4 <- c(5.8, 5.3, 3.2, 7.5)
>
> nematodes <- data.frame(value = c(group1, group2, group3, group4),
+ group = factor(rep(c("Group 1", "Group 2", "Group 3", "Group 4"), each = 4)))
>
> # --------------------(a)----------------------
> # Conduct the ANOVA
> fit <- aov(value ~ group, data = nematodes)
> summary(fit)
            Df Sum Sq Mean Sq F value   Pr(>F)
group        3 100.65   33.55   12.08 0.000616 ***
Residuals   12  33.33    2.78
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output of the ANOVA shows that the p-value is 0.000616, which is less than the significance level of 0.05. Therefore, we reject the null hypothesis that the mean growth is the same across all nematode treatment groups. In other words, there is strong evidence that the nematodes have an effect on the plant growth.

(b) The appropriate contrast to test whether the introduction of nematodes reduces plant growth would be to compare the mean plant growth in the nematode treatment group to the mean plant growth in the control group. To set up this contrast, we can use weights of -1 for the control group and 1/3 for each of the three nematode treatment groups. $weights < -c(-1, 1/3, 1/3, 1/3)$

```
> # --------------------(b)----------------------
> # Define the contrast weights
> weights <- c(-1, 1/3, 1/3, 1/3)
> with_nematodes <- c(weights[2]*group2, weights[3]*group3, weights[4]*group4)
> without_nematodes <- weights[1]*group1
> t.test(without_nematodes, with_nematodes)

        Welch Two Sample t-test

data:  without_nematodes and with_nematodes
t = -12.288, df = 3.4155, p-value = 0.0006028
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -16.191488  -9.880734
sample estimates:
 mean of x  mean of y
-10.650000   2.386111
```

The p-value is 0.0006028 which is less than 0.05, so we reject the null hypothesis that the mean plant growth in the nematode treatment group is significantly different from the mean plant growth in the control group and conclude that introduction of nematodes changes plant growth.

(c) To test whether the groups treated with 1000 nematodes per plant and 5000 nematodes per plant have different plant growth, we can use the following contrast:

$$weights < -c(0, -1, 1, 0)$$

This contrast will compare the average growth of the group treated with 5000 nematodes per plant to the average growth of the group treated with 1000 nematodes per plant, while ignoring the control group and the group treated with 10000 nematodes per plant.

```
> # --------------------(c)----------------------
> # Define the contrast weights
> weights2 <- c(0, -1, 1, 0)
>
> nematodes_1000 <- weights2[2]*group2
> nematodes_5000 <- weights2[3]*group3
>
> t.test(nematodes_1000, nematodes_5000)

        Welch Two Sample t-test

data:  nematodes_1000 and nematodes_5000
t = -16.538, df = 5.8189, p-value = 4.1e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -18.41405 -13.63595
sample estimates:
mean of x mean of y
  -10.425     5.600
```

The *p-value* for the contrast is 0.0000041, which is less than 0.05 significance level. Therefore, we have enough evidence to reject the null hypothesis that there is no difference in plant growth between the groups treated with 1000 and 5000 nematodes per plant. In other words, we can statistically claim that these two groups have different plant growth.

5. (40 points)

A study was carried out to see how previous work experience improves the ability of a programmer to complete a complex programming task. 26 randomly chosen programmers were asked to complete a complex programming task within a specified time period. For each programmer, the length of previous programming work experience (in months) was recorded as well as whether the task was successfully completed (1 indicate success). The data was presented in the following table.

(a) Does work experience improve the programmer's ability? Carry out appropriate statistical analysis. What is your conclusion?

(b) With 95% confidence, give an estimation interval of the improvement in the odds of completing the task with specified time period for each extra year of work experience. (Note the data has unit of month.)

(c) The employer values the employee according to the probability of finishing the given task within the time period. A person with 100% probability of finishing the given task within the time period is worth the pay of $90,000 per year; A person with 80% probability of finishing is worth $72,000, etc. Knowing only the work history of an applicant, what salary (X dollars per year) should the employer pay a programmer with 24 months of previous work experience?

(d) A second programmer with 18 months of work experience is willing to accept a yearly salary of $(X-10,000), where X is the answer that you got in part (c) above. According to this analysis, is the second programmer a better deal for the company than the first programmer (salaried at $X) in part (c)?

Experience in months	*Success*
14	**0**
29	**0**
6	**0**
25	**1**
18	**1**
4	**0**
18	**0**
12	**0**
22	**1**
6	**0**
30	**1**
11	**0**
30	**1**
5	**0**
20	**1**
13	**0**
9	**0**
32	**1**
24	**0**
13	**1**
19	**0**
4	**0**
28	**1**
22	**1**
8	**1**
14	**0**

Solution: (a)

```
> # --------------------(a)----------------------
> model <- glm(success ~ Experience, data = programmer, family = binomial)
> summary(model)

Call:
glm(formula = success ~ Experience, family = binomial, data = programmer)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8924  -0.7591  -0.4030   0.7715   2.0147

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.2206     1.2770  -2.522   0.0117 *
Experience    0.1665     0.0659   2.527   0.0115 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35.426  on 25  degrees of freedom
Residual deviance: 26.140  on 24  degrees of freedom
AIC: 30.14

Number of Fisher Scoring iterations: 4
```

The output shows that the coefficient of the experience variable is 0.1665, and the p-value of the coefficient is 0.0115. The p-value is smaller than 0.05, which means that we can reject the null hypothesis that there is no relationship between work experience and the ability to complete a complex programming task. Therefore, we can conclude that work experience improves the programmer's ability.

(b)

```
> # --------------------(b)----------------------
> confint(model, level = 0.95)
Waiting for profiling to be done...
                 2.5 %     97.5 %
(Intercept) -6.23412075 -1.0523797
Experience   0.05370693  0.3212723
```

The 95% confidence interval for the coefficient of the variable "Experience" is (0.0537, 0.3213).

But since the data has the unit of month, this means that with 95% confidence, for every extra year of work experience, the odds of completing the task within the specified time period increase by a factor between exp(0.0537*12) = 1.904844 and exp(0.3213*12) = 47.25696.

(c) To calculate the salary of a programmer with 24 months of previous experience, we need to first estimate the probability of successfully completing the task within the specified time period. Then we can use the given valuation scheme to determine the appropriate salary.

We can use the logistic regression model that we fitted in (a) to estimate the probability of success for a programmer with 24 months of previous work experience. We simply plug in the value of 24 for Experience in the model equation:

```
> # --------------------(b)----------------------
> probability <- predict(model, newdata = data.frame(Experience = 24), type = "response")
> probability
        1
0.6847214
```

So, the estimated probability of success for a programmer with 24 months of previous work experience is 0.6847214.

Using the given valuation scheme, we can determine the appropriate salary for this programmer as follows:

- If the probability of success is 100%, the salary is $90,000 per year.
- If the probability of success is 80%, the salary is $72,000 per year.
- For a linearly decreasing scale of probability, the salary can be calculated using the formula: $salary = probability * \$90{,}000$

Therefore, for a programmer with 24 months of previous work experience, the appropriate salary is:

```
> salary <- ifelse(probability == 1, 90000,
+           ifelse(probability >= 0.8, 72000,
+           (probability * 90000)))
> salary
       1
61624.92
```

So, the appropriate salary for a programmer with 24 months of previous work experience is approximately $61624.92 per year.

(d) According to part (c), a programmer with 24 months of work experience should be paid $61624.92 per year to ensure a 68.5% probability of completing a given task within a specified time period.

Now, the second programmer with 18 months of work experience is willing to accept a yearly salary of $(X-10,000), where X is $61624.92 from part (c), i.e., $51624.92.

To compare the two programmers, we need to calculate the probability of completing the task within the specified time period for the second programmer with 18 months of work experience. Using the logistic regression model, we get:

```
> # --------------------(d)----------------------
> prob18 <- predict(model, newdata = data.frame(Experience = 18))
> time <- exp(prob18) / (1 + exp(prob18))
> time
        1
0.4443621
```

The second programmer has a probability of approximately 44.44% of completing the task within the specified time-period.

Now, we need to compare the two probabilities of task completion and salaries. The first programmer has a probability of 0.6847 of completing the task within the specified time-period, whereas the second programmer has a probability of only 0.4444 of completing the same task. Moreover, the second programmer is expecting a salary of $51624.92 per year, which is greater than what he deserves ($39,992.59) based on his probability of completing task. Therefore, based on this analysis, it can be

concluded that the first programmer with 24 months of work experience is a better deal for the company than the second programmer with 18 months of work experience.

6. (20 points) Do 20.6.7 on page 506 of the textbook

Ex. 20.6.7. In the 1980s, a study was conducted to examine the effects of the drug ganciclovir on HIV/AIDS patients suffering from disseminated cytomegalovirus infection [303]. Two groups of patients were followed; 18 were treated with the drug, and 11 were not. The results of this study are contained in the dataset *cytomegalo*. Survival times in months after diagnosis are saved under the variable name time, and indicators of censoring status where 1 designates that a death occurred and 0 that an observation was censored - under the name death. Values of treatment group, where 1 indicates that a patient took the drug and 0 that he or she did not, are saved under the name group.

(a) How many deaths occurred in each treatment group?

(b) Use the product-limit method to estimate the survival function for each treatment group.

(c) Construct survival curves for the two treatment groups based on the product-limit estimate of S(1).

(d) Does it appear that the individuals in one group survive longer than those in the other group?

(e) Use the log-rank test to evaluate the null hypothesis that the distributions of survival times are identical in the two groups. What do you conclude?

Solution: (a)

```
> # --------------------(a)----------------------
> table(cytomegalo$group, cytomegalo$death)

    0  1
0   0 11
1   3 15
```

Therefore, there were 15 deaths in the treatment group and 11 deaths in the non-treatment group.

(b) To estimate the survival function using the product-limit method, we can use the survfit() function from the survival package in R:

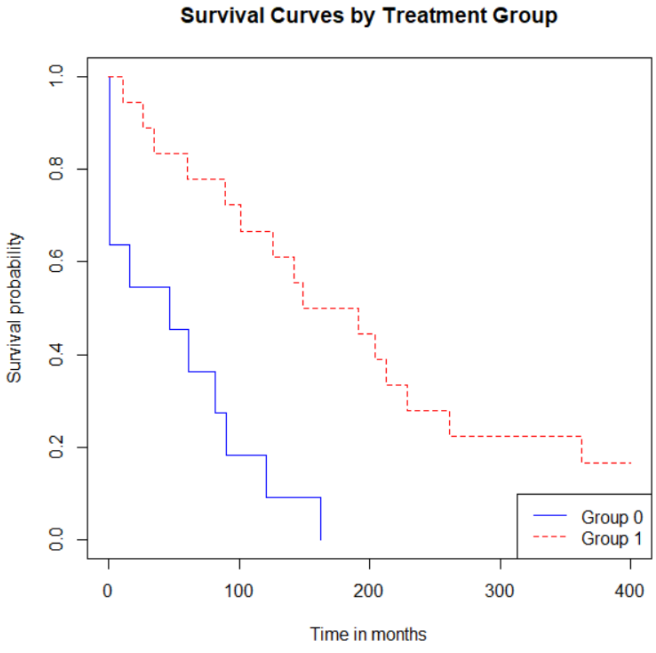
```
> # ----------------------(b)-----------------------
> library(survival)
>
> fit <- survfit(Surv(time, death) ~ group, data = cytomegalo)
> summary(fit)
Call: survfit(formula = Surv(time, death) ~ group, data = cytomegalo)

                group=0
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     11       4   0.6364  0.1450       0.4071        0.995
   16      7       1   0.5455  0.1501       0.3180        0.936
   47      6       1   0.4545  0.1501       0.2379        0.868
   61      5       1   0.3636  0.1450       0.1664        0.795
   82      4       1   0.2727  0.1343       0.1039        0.716
   90      3       1   0.1818  0.1163       0.0519        0.637
  121      2       1   0.0909  0.0867       0.0140        0.589
  162      1       1   0.0000     NaN           NA           NA

                group=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   11     18       1    0.944  0.0540       0.8443        1.000
   26     17       1    0.889  0.0741       0.7549        1.000
   35     16       1    0.833  0.0878       0.6778        1.000
   60     15       1    0.778  0.0980       0.6076        0.996
   89     14       1    0.722  0.1056       0.5423        0.962
  101     13       1    0.667  0.1111       0.4809        0.924
  126     12       1    0.611  0.1149       0.4227        0.883
  142     11       1    0.556  0.1171       0.3675        0.840
  149     10       1    0.500  0.1179       0.3150        0.794
  191      9       1    0.444  0.1171       0.2652        0.745
  204      8       1    0.389  0.1149       0.2179        0.694
  213      7       1    0.333  0.1111       0.1734        0.641
  229      6       1    0.278  0.1056       0.1319        0.585
  261      5       1    0.222  0.0980       0.0936        0.527
  362      4       1    0.167  0.0878       0.0593        0.468
```

The output shows the survival function estimates for each treatment group at each observed time point. For example, for the non-treatment group, the estimated survival probability at time 1 months is 0.6364 with a standard error of 0.1450, and the estimated survival probability at time 16 months is 0.5455 with a standard error of 0.1501, and so on.

(c)



(d) Yes, it seems that those in group 1 (those who had ganciclovir treatment) live longer than those in group 0 (those who did not get treatment). The survival curves may be observed in the figure, where group 1's curve is continuously above group 0's curve. To assess the significance of this difference, though, a proper statistical test is required.

(e)

```
> # --------------------(e)----------------------
> survdiff(Surv(time, death) ~ group, data = cytomegalo)
Call:
survdiff(formula = Surv(time, death) ~ group, data = cytomegalo)

          N Observed Expected (O-E)^2/E (O-E)^2/V
group=0 11       11     4.63      8.74      12.4
group=1 18       15    21.37      1.90      12.4

 Chisq= 12.4  on 1 degrees of freedom, p= 4e-04
```

The log-rank test evaluates whether there is a significant difference between the survival curves of the two groups.

The null hypothesis is that the survival curves are identical in the two groups.

The alternative hypothesis is that the survival curves differ between the two groups.

The test shows that the chi-squared statistic is 12.4 with 1 degree of freedom and a very small p-value of 4e-04. Since the p-value is less than 0.05, we reject the null hypothesis and conclude that there is a statistically significant difference in survival between the two groups.

7. For the following data, we wish to test if X has the same mean in group A and group B.

(a) (3 points) Conduct the two sample t-test for this hypothesis.

(b) (7 points) Please conduct a *permutation test* using the t-test statistic in part (a). Submit your R code for this permutation test, the R outputs and results of your test. (Hint: while you can mimic the example R code for the permutation correlation test, please note that the way of permutation is different. See the lecture notes at end of the non-parametric test module.)

The data is also in the file PermutationTestData.txt

	X	GROUP
1	9.02	A
2	7.88	A
3	10.80	A
4	7.97	A
5	7.81	A
6	8.91	A
7	8.43	A
8	8.88	A
9	8.91	A
10	9.42	A
11	13.16	A
12	9.05	A

13	7.32	B
14	7.95	B
15	7.32	B
16	8.61	B
17	7.01	B
18	7.09	B
19	9.17	B
20	7.78	B
21	7.51	B
22	10.92	B
23	8.09	B
24	7.07	B
25	7.02	B
26	8.46	B
27	9.23	B

Solution: (a) Assuming that Variance of X in both groups are equal.

```
> # --------------------(a)----------------------
> # Conduct a two-sample t-test
> t.test(X ~ Group, data = data, mu = 0, var.equal = TRUE)

        Two Sample t-test

data:  X by Group
t = 2.3123, df = 25, p-value = 0.02928
alternative hypothesis: true difference in means between group A and group B is not equal to 0
95 percent confidence interval:
 0.1257304 2.1742696
sample estimates:
mean in group A mean in group B
       9.186667        8.036667
```

The two-sample t-test indicates that the t-statistic is 2.3123 and the p-value is 0.02928.

The alternative hypothesis is that the true difference in means between group A and group B is not equal to 0. The 95% confidence interval for the difference in means is (0.1257304, 2.1742696), which indicates that the mean of X in group A is likely to be higher than that in group B. The sample mean for group A is 9.186667 and for group B is 8.036667.

The results suggest that there is evidence to reject the null hypothesis that the means of X in group A and group B are equal and conclude that the mean of X in group A and group B are not same.

(b)

```
> # --------------------(b)----------------------
> # Define a function to calculate the two-sample t-test statistic
> t_statistic <- function(x, group) {
+   mean_diff <- abs(diff(tapply(x, group, mean)))
+   se <- sqrt(sum(tapply(x, group, var) / table(group)))
+   t_val <- mean_diff / se
+   return(t_val)
+ }
>
> # Calculate the observed t-test statistic
> obs_t <- t.test(X ~ Group, data = data, mu = 0, var.equal = TRUE)$statistic
>
> # Set the number of permutations
> n_perm <- 10000
>
> # Create a vector to store permuted t-test statistics
> perm_t <- rep(NA, n_perm)
>
> # Permute the group labels and calculate t-test statistics for each permutation
> for (i in 1:n_perm) {
+   perm_group <- sample(data$Group)
+   perm_t[i] <- t_statistic(data$X, perm_group)
+ }
>
> # Calculate the p-value as the proportion of permuted t-test statistics that are as extreme
> # as the observed t-test statistic in the direction of the alternative hypothesis
> p_val <- mean(abs(perm_t) >= abs(obs_t))
>
> # Print the results
> cat("Observed t-test statistic:", obs_t, "\n")
Observed t-test statistic: 2.312349
> cat("Permutation test p-value:", p_val, "\n")
Permutation test p-value: 0.0251
>
```

This permutation test indicates that there is a significant difference between group A and group B in the mean of X, and we would reject the null hypothesis based on this finding at the 5% level of significance.

The parametric t-test's p-value (0.02928) and the permutation test's p-value (0.0251) are quite comparable, and both tests come to the same result that the mean of X in group A and group B are not same.