Homework 4 Submission

Yahriel Salinas-Reyes

2024-10-28

# Question 1

1. Evaluate the null hypothesis that the population proportions of students who drove while drinking are the same in the two calendar years (2010 and 2014).

For this analysis, we can use the Chi-square test to compare the proportions of students who drove while drinking in both years.

# Create the contingency table  
data1 <- matrix(c(1250, 991, 1387, 1666), nrow = 2, byrow = TRUE)  
colnames(data1) <- c("2010", "2014")  
rownames(data1) <- c("Yes", "No")  
data1 <- as.table(data1)  
  
# Perform the Chi-square test  
chi\_square\_test <- chisq.test(data1)  
  
# Print the result  
print(chi\_square\_test)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: data1  
## X-squared = 54.942, df = 1, p-value = 1.241e-13

1. What do you conclude about the behavior of college students?

Since the p-value (1.241e-13) from the Chi-squared rest is less than the significance level, we reject the null hypothesis, indicating that there is a significant difference in the proportions of students who drove while drinking between the two years.

# Question 2

1. Do these data support the null hypothesis that there is no association between the time of screening and diagnosis?

Here, we can also use the Chi-square test to assess the association between the first and second screenings.

# Create the contingency table  
data2 <- matrix(c(1763, 489, 403, 670), nrow = 2, byrow = TRUE)  
colnames(data2) <- c("Present", "Absent")  
rownames(data2) <- c("Present", "Absent")  
data2 <- as.table(data2)  
  
# Perform the Chi-square test  
chi\_square\_test2 <- chisq.test(data2)  
  
# Print the result  
print(chi\_square\_test2)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: data2  
## X-squared = 529.09, df = 1, p-value < 2.2e-16

Since the p-value (< 2.2e-16) from the Chi-squared rest is less than the significance level, we reject the null hypothesis, indicating that there is a significant difference or association between the time of screening and diagnosis.

1. The data could also be displayed in a different manner. Is there anything wrong with this presentation? Why?

The alternative presentation aggregates the results, which could obscure variability in the data. This format might hide differences in screening consistency between the two tests, making it challenging to assess intra-observer reliability accurately.

# Question 3

1. Test the null hypothesis that there is no association between drinking status on the two different types of questionnaires.

Since we want to compare two categorical variables (drinking status from two different questionnaires), a Chi-square test is appropriate.

# Load the necessary library  
library(readxl)  
  
# Import the dataset  
alcohol\_data <- read\_excel("alcohol.xls")  
  
# Create a contingency table  
contingency\_table <- table(alcohol\_data$genques, alcohol\_data$alcques)  
  
# Perform the Chi-square test  
chi\_square\_result <- chisq.test(contingency\_table)  
  
# Display the result  
print(chi\_square\_result)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: contingency\_table  
## X-squared = 103.21, df = 1, p-value < 2.2e-16

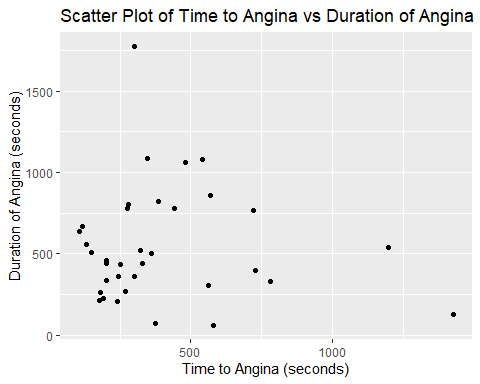
1. What do you conclude?

Based on the p-value from the Chi-square test result which is less than 0.05, we reject the null hypothesis and we can conclude there is a significant association between the two questionnaires.

# Question 4

1. Create a two-way scatter plot for these data.

# Load the necessary library  
library(readxl)  
library(ggplot2)  
  
# Import the dataset  
ischemic\_data <- read\_excel("ischemic.xls")  
  
# Create a scatter plot  
ggplot(ischemic\_data, aes(x = time, y = duration)) +  
 geom\_point() +  
 labs(title = "Scatter Plot of Time to Angina vs Duration of Angina",  
 x = "Time to Angina (seconds)",  
 y = "Duration of Angina (seconds)")

 b. In the population of patients with ischemic heart disease, does there appear to be any evidence of a linear relationship between time to angina and the duration of the attack?

We can inspect the scatter plot for a linear trend. Since the points suggest a linear relationship, we can proceed with correlation analysis.

1. In calculating the correlation coefficient, which test will you use, Pearson or Spearman? Why?

Since both variables are continuous and likely normally distributed, use Pearson’s correlation. If not normally distributed, use Spearman.

1. Test the null hypothesis that the population correlation is equal to 0.

# Perform hypothesis test for correlation  
cor\_test\_result <- cor.test(ischemic\_data$time, ischemic\_data$duration)  
  
# Print results  
cor\_test\_result

##   
## Pearson's product-moment correlation  
##   
## data: ischemic\_data$time and ischemic\_data$duration  
## t = -0.42465, df = 33, p-value = 0.6738  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.3972091 0.2660620  
## sample estimates:  
## cor   
## -0.07372094

Since the p-value here is greater than 0.05, then we fail to reject the null and conclude that there is no significant correlation, correlation = 0.

# Question 5

Use Spearman’s rank correlation because the Apgar score is ordinal.

# Import dataset  
lowbwt <- read\_excel("lowbwt.xls")  
  
  
  
# Conduct Spearman correlation  
spearman\_result <- cor.test(lowbwt$sbp, lowbwt$apgar5, method="spearman")

## Warning in cor.test.default(lowbwt$sbp, lowbwt$apgar5, method = "spearman"):  
## Cannot compute exact p-value with ties

# Print results  
spearman\_result

##   
## Spearman's rank correlation rho  
##   
## data: lowbwt$sbp and lowbwt$apgar5  
## S = 148593, p-value = 0.2832  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 0.1083551

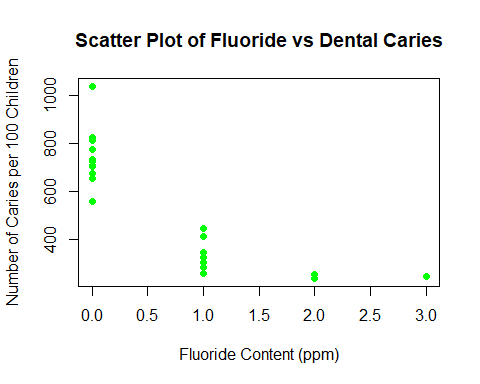
1. Null hypothesis and conclusion?

We fail to reject the null hypothesis that there is no correlation between systolic blood pressure and Apgar score, since the p-value is greater than 0.05.

# Question 6

1. Construct a scatter plot for these data.
2. Calculate correlation.
3. Is this correlation significantly different from 0?

# Import dataset  
water <- read\_excel("water.xls")  
  
# Create scatter plot  
plot(water$fluoride, water$caries,  
 main="Scatter Plot of Fluoride vs Dental Caries",  
 xlab="Fluoride Content (ppm)",  
 ylab="Number of Caries per 100 Children",  
 pch=19, col='green')



## B.   
# Calculate correlation  
fluoride\_caries\_correlation <- cor(water$fluoride, water$caries)  
fluoride\_caries\_correlation

## [1] -0.8265237

## C.   
# Conduct hypothesis test  
correlation\_test <- cor.test(water$fluoride, water$caries)  
  
# Print results  
correlation\_test

##   
## Pearson's product-moment correlation  
##   
## data: water$fluoride and water$caries  
## t = -6.4003, df = 19, p-value = 3.878e-06  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.9273372 -0.6138627  
## sample estimates:  
## cor   
## -0.8265237

Since the p-value is than 0.05, we conclude that the correlation is significantly different from 0.

# Question 7

1. Calculate correlations between ranks.
2. Is each correlation significantly different from 0?

# Import dataset  
actions <- read\_excel("actions.xls")  
  
# Calculate correlations  
correlation\_1991\_1992 <- cor(actions$rank91, actions$rank92)  
correlation\_1991\_1993 <- cor(actions$rank91, actions$rank93)  
correlation\_1991\_1994 <- cor(actions$rank91, actions$rank94)  
correlation\_1991\_1995 <- cor(actions$rank91, actions$rank95)  
  
# Print results  
correlation\_1991\_1992

## [1] 0.810236

correlation\_1991\_1993

## [1] 0.764272

correlation\_1991\_1994

## [1] 0.6292114

correlation\_1991\_1995

## [1] 0.5735813

## The Correlation decreases with the increasing year.   
  
## B.   
# Conduct hypothesis tests  
test\_1991\_1992 <- cor.test(actions$rank91, actions$rank92)  
test\_1991\_1993 <- cor.test(actions$rank91, actions$rank93)  
test\_1991\_1994 <- cor.test(actions$rank91, actions$rank94)  
test\_1991\_1995 <- cor.test(actions$rank91, actions$rank95)  
  
# Print results  
test\_1991\_1992

##   
## Pearson's product-moment correlation  
##   
## data: actions$rank91 and actions$rank92  
## t = 9.6769, df = 49, p-value = 5.921e-13  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.6883532 0.8876240  
## sample estimates:  
## cor   
## 0.810236

test\_1991\_1993

##   
## Pearson's product-moment correlation  
##   
## data: actions$rank91 and actions$rank93  
## t = 8.2958, df = 49, p-value = 6.768e-11  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.6190793 0.8589441  
## sample estimates:  
## cor   
## 0.764272

test\_1991\_1994

##   
## Pearson's product-moment correlation  
##   
## data: actions$rank91 and actions$rank94  
## t = 5.6669, df = 49, p-value = 7.598e-07  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.4278102 0.7710881  
## sample estimates:  
## cor   
## 0.6292114

test\_1991\_1995

##   
## Pearson's product-moment correlation  
##   
## data: actions$rank91 and actions$rank95  
## t = 4.9015, df = 49, p-value = 1.084e-05  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.3539457 0.7332590  
## sample estimates:  
## cor   
## 0.5735813

## Yes, they all are significanctly different from zero.

1. Are states equally strict?

Since the correlations are low as years increase, it may indicate inconsistency among states.

# Question 8

# Import dataset  
library(readxl)  
lowbwt <- read\_excel("lowbwt.xls")  
  
  
# a. Compute least-squares regression line  
lm\_sbp\_gestage <- lm(sbp ~ gestage, data = lowbwt)  
summary(lm\_sbp\_gestage)

##   
## Call:  
## lm(formula = sbp ~ gestage, data = lowbwt)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23.162 -7.828 -1.483 5.568 39.781   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.5521 12.6506 0.834 0.40625   
## gestage 1.2644 0.4362 2.898 0.00463 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11 on 98 degrees of freedom  
## Multiple R-squared: 0.07895, Adjusted R-squared: 0.06956   
## F-statistic: 8.401 on 1 and 98 DF, p-value: 0.004628

# Interpretation: The estimated slope (b1) represents the change in systolic blood pressure for each week increase in gestational age.  
  
# b. Hypothesis Test for Slope  
summary(lm\_sbp\_gestage)$coefficients

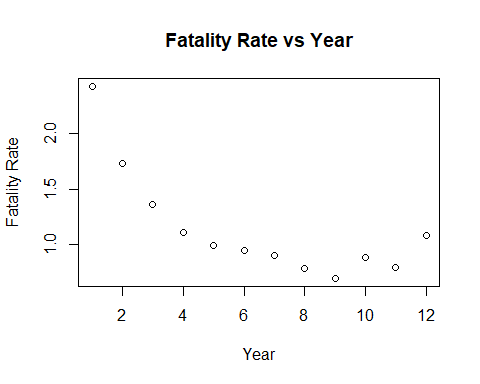
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 10.55207 12.6506265 0.8341143 0.406245131  
## gestage 1.26438 0.4362311 2.8984173 0.004628003

# Interpretation: Since the P-value for b1 is less than 0.05, we conclude that b1 differs significantly from zero.  
  
# c. Predicted systolic blood pressure at gestational age of 31 weeks  
predict(lm\_sbp\_gestage, newdata = data.frame(gestage = 31))

## 1   
## 49.74784

# Question 9

# Load dataset  
miner <- data.frame(year=1:12, rate=c(2.419, 1.732, 1.361, 1.108, 0.996, 0.952, 0.904, 0.792, 0.701, 0.890, 0.799, 1.084))  
  
# a) Scatter plot  
plot(miner$year, miner$rate, main="Fatality Rate vs Year", xlab="Year", ylab="Fatality Rate")



# b) Least-squares regression line  
model\_9b <- lm(rate ~ year, data=miner)  
summary(model\_9b)

##   
## Call:  
## lm(formula = rate ~ year, data = miner)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.3013 -0.2112 -0.1647 0.1162 0.7150   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.80561 0.21012 8.593 6.26e-06 \*\*\*  
## year -0.10166 0.02855 -3.561 0.00518 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3414 on 10 degrees of freedom  
## Multiple R-squared: 0.559, Adjusted R-squared: 0.515   
## F-statistic: 12.68 on 1 and 10 DF, p-value: 0.005176

# c) Transformation with ln(x)  
miner$log\_year <- log(miner$year)  
model\_9c <- lm(rate ~ log\_year, data=miner)  
summary(model\_9c)

##   
## Call:  
## lm(formula = rate ~ log\_year, data = miner)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.20290 -0.12265 -0.09047 0.09815 0.42630   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.13515 0.14879 14.350 5.35e-08 \*\*\*  
## log\_year -0.59457 0.08193 -7.257 2.74e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2054 on 10 degrees of freedom  
## Multiple R-squared: 0.8404, Adjusted R-squared: 0.8244   
## F-statistic: 52.66 on 1 and 10 DF, p-value: 2.737e-05

# d) Transformation with 1/x  
miner$inv\_year <- 1 / miner$year  
model\_9d <- lm(rate ~ inv\_year, data=miner)  
summary(model\_9d)

##   
## Call:  
## lm(formula = rate ~ inv\_year, data = miner)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.17711 -0.04879 -0.02905 0.04424 0.25612   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.67718 0.05063 13.38 1.05e-07 \*\*\*  
## inv\_year 1.80840 0.14019 12.90 1.48e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1224 on 10 degrees of freedom  
## Multiple R-squared: 0.9433, Adjusted R-squared: 0.9376   
## F-statistic: 166.4 on 1 and 10 DF, p-value: 1.476e-07

# e) Choose best model based on R-squared and p-value comparisons  
## Based on the R^2 value and p-values, the first model had the moderate R^2 with the at least half of the variance explained and the p-value just above the significance interval suggesting near statistical significance.  
  
# f) Use best model to predict rate for year 9  
predict(model\_9b, newdata=data.frame(year=9)) # Replace model if another is best

## 1   
## 0.89069

# Question 10

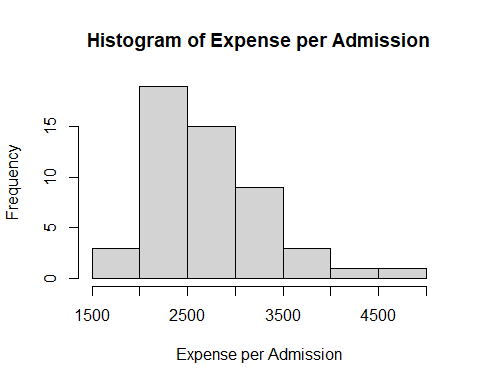
# a) Numerical summaries  
summary(hospital$expadm)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1772 2260 2600 2717 3054 4612

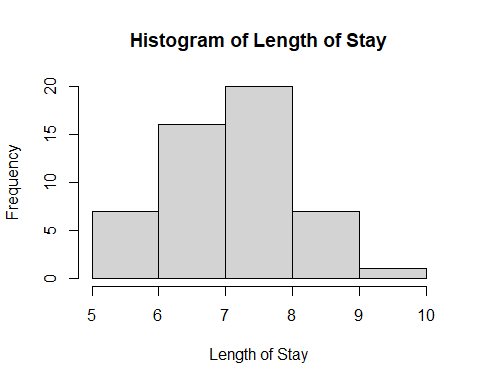
summary(hospital$los)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 5.000 7.000 8.000 7.569 8.000 10.000

# b) Histograms  
hist(hospital$expadm, main = "Histogram of Expense per Admission", xlab = "Expense per Admission")



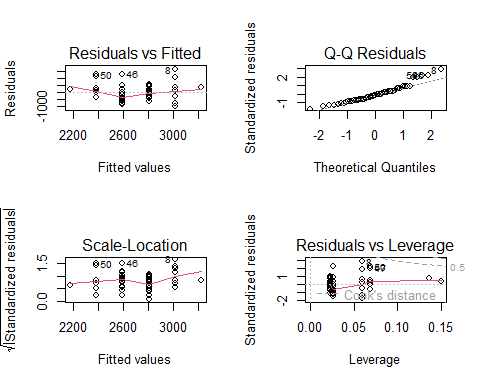
hist(hospital$los, main = "Histogram of Length of Stay", xlab = "Length of Stay")



# c) Least-squares regression line  
model\_exp\_los <- lm(expadm ~ los, data = hospital)  
summary(model\_exp\_los)

##   
## Call:  
## lm(formula = expadm ~ los, data = hospital)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -957.30 -400.70 -83.06 238.04 1595.70   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1133.14 614.28 1.845 0.0711 .  
## los 209.24 80.47 2.600 0.0123 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 571.9 on 49 degrees of freedom  
## Multiple R-squared: 0.1213, Adjusted R-squared: 0.1033   
## F-statistic: 6.761 on 1 and 49 DF, p-value: 0.01228

# d) Interpretation of the slope  
# This describes the change in expense per admission with each additional day in length of stay. While the p-value for the b1 is less than .05, we can conclude that b1 differs significancly from zero.  
  
# e) Diagnostic plots (4 plots)  
par(mfrow=c(2,2))  
plot(model\_exp\_los)



# Question 11

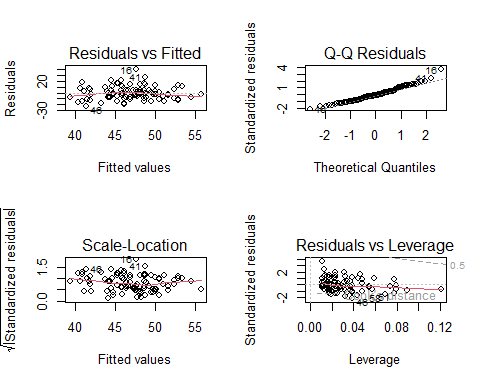
# Load dataset  
lowbwt <- read\_excel("lowbwt.xls")  
  
# (a) Construct a model using gestational age and Apgar score to predict systolic blood pressure  
model\_11 <- lm(sbp ~ gestage + apgar5, data = lowbwt)  
summary(model\_11)

##   
## Call:  
## lm(formula = sbp ~ gestage + apgar5, data = lowbwt)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.374 -8.180 -1.088 4.985 39.424   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.8034 12.6629 0.774 0.4407   
## gestage 1.1848 0.4424 2.678 0.0087 \*\*  
## apgar5 0.4875 0.4613 1.057 0.2932   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.99 on 97 degrees of freedom  
## Multiple R-squared: 0.08944, Adjusted R-squared: 0.07066   
## F-statistic: 4.764 on 2 and 97 DF, p-value: 0.01063

# (b) Test the null hypothesis that b1 = 0  
# When we check the p-value for gestage, we see that it is below the significance level, just like the overall p-value from the summary. We reject the null hypothesis suggesting statistical significance.  
  
# (c) Interpretation of b1 and b2 (gestage and apgar5 coefficients)  
# From the summary, we can see that both b1 and b2 are positive values for positive correlations, for b1 it is significant with a p-value below .05, while b2 is not with a p-value over the significance level.   
  
  
# (d) Predicted systolic BP for gestage = 31 and apgar5 = 7  
predict(model\_11, data.frame(gestage = 31, apgar5 = 7))

## 1   
## 49.94562

# (e) Interpret R-squared  
# From the summary output, we see a very low R^2 value, suggesting that our model has a low amount of explained variance. While error is still present, it is not explained by the model.  
  
# (f) Assumption check (normality of residuals)  
par(mfrow = c(2, 2))  
plot(model\_11)



# Question 12

# (a) Construct model with gestage, apgar5, and sex as predictors  
model\_12a <- lm(sbp ~ gestage + apgar5 + sex, data = lowbwt)  
summary(model\_12a)

##   
## Call:  
## lm(formula = sbp ~ gestage + apgar5 + sex, data = lowbwt)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.854 -7.996 -1.046 5.092 38.737   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.3235 12.7371 0.732 0.46595   
## gestage 1.1853 0.4440 2.670 0.00892 \*\*  
## apgar5 0.4747 0.4635 1.024 0.30840   
## sex 1.2434 2.2253 0.559 0.57764   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11.03 on 96 degrees of freedom  
## Multiple R-squared: 0.09239, Adjusted R-squared: 0.06403   
## F-statistic: 3.257 on 3 and 96 DF, p-value: 0.02494

# (b) Adding interaction between gestage and sex  
model\_12b <- lm(sbp ~ gestage \* sex + apgar5, data = lowbwt)  
summary(model\_12b)

##   
## Call:  
## lm(formula = sbp ~ gestage \* sex + apgar5, data = lowbwt)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.534 -7.660 -0.943 5.246 38.697   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 14.4355 15.2461 0.947 0.3461   
## gestage 1.0072 0.5313 1.896 0.0611 .  
## sex -15.7541 27.7402 -0.568 0.5714   
## apgar5 0.4795 0.4651 1.031 0.3052   
## gestage:sex 0.5881 0.9567 0.615 0.5402   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11.07 on 95 degrees of freedom  
## Multiple R-squared: 0.09599, Adjusted R-squared: 0.05792   
## F-statistic: 2.522 on 4 and 95 DF, p-value: 0.04606

# Based on the P-values which for both models are below .05, they both have statistical significance. The first one has a lower score. But looking at the interaction term in the 2nd model, it's p-value is quite high suggesting that the interaction doesn't have statistical significance.

# Question 13

# Load dataset  
heart <- read\_excel("heart.xls")  
  
  
# Fit separate linear regression models for PDI and MDI  
model\_13\_pdi <- lm(pdi ~ trtment, data = heart)  
summary(model\_13\_pdi)

##   
## Call:  
## lm(formula = pdi ~ trtment, data = heart)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -47.771 -6.845 0.229 12.229 42.082   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 91.918 1.830 50.240 <2e-16 \*\*\*  
## trtment 5.854 2.615 2.239 0.0268 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.63 on 141 degrees of freedom  
## (14 observations deleted due to missingness)  
## Multiple R-squared: 0.03432, Adjusted R-squared: 0.02747   
## F-statistic: 5.011 on 1 and 141 DF, p-value: 0.02676

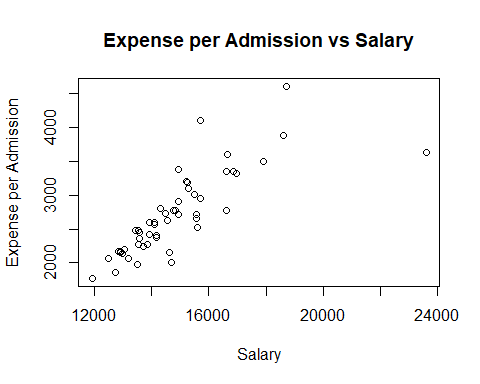
model\_13\_mdi <- lm(mdi ~ trtment, data = heart)  
summary(model\_13\_mdi)

##   
## Call:  
## lm(formula = mdi ~ trtment, data = heart)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -56.400 -9.222 0.600 11.028 38.838   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 103.162 1.810 56.981 <2e-16 \*\*\*  
## trtment 3.238 2.597 1.247 0.214   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.57 on 142 degrees of freedom  
## (13 observations deleted due to missingness)  
## Multiple R-squared: 0.01083, Adjusted R-squared: 0.003865   
## F-statistic: 1.555 on 1 and 142 DF, p-value: 0.2145

# The first model with pdi was the more statistically significant model with a p-value below the significance level.

# Question 14

# a. Scatter plot  
plot(hospital$salary, hospital$expadm, main = "Expense per Admission vs Salary", xlab = "Salary", ylab = "Expense per Admission")



# b. Multiple regression  
model9 <- lm(expadm ~ los + salary, data = hospital)  
summary(model9)

##   
## Call:  
## lm(formula = expadm ~ los + salary, data = hospital)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -896.91 -112.25 -8.92 100.78 854.05   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.686e+03 4.537e+02 -5.920 3.32e-07 \*\*\*  
## los 2.269e+02 4.140e+01 5.482 1.53e-06 \*\*\*  
## salary 2.481e-01 2.117e-02 11.722 1.09e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 294 on 48 degrees of freedom  
## Multiple R-squared: 0.7725, Adjusted R-squared: 0.763   
## F-statistic: 81.5 on 2 and 48 DF, p-value: 3.693e-16

# c.   
# Yes, the addition of length of stay does make the model better. From the summary we can see that the p-values for b1 and overall are well beloe 0.05, suggesting they have a high significance value. The R^2 value is also high, suggesting a high amount of variance in the model can be explained.

# Question 15

detroit <- read\_excel("detroit.xls")  
# a. Multiple linear regression for homicide with predictors  
lm\_homicide <- lm(homicide ~ register + police + unemp + weekly, data = detroit)  
summary(lm\_homicide)

##   
## Call:  
## lm(formula = homicide ~ register + police + unemp + weekly, data = detroit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.714 -1.213 -0.008 2.036 4.135   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -54.249809 11.022760 -4.922 0.00116 \*\*  
## register 0.016960 0.005277 3.214 0.01235 \*   
## police 0.176436 0.084858 2.079 0.07120 .   
## unemp 0.255979 0.538099 0.476 0.64700   
## weekly 0.087424 0.073741 1.186 0.26982   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.276 on 8 degrees of freedom  
## Multiple R-squared: 0.9725, Adjusted R-squared: 0.9588   
## F-statistic: 70.84 on 4 and 8 DF, p-value: 2.78e-06

#b-d  
# Based on the summary, register is the only variable which has a significant effect (but low positive correlation) on homocide rate. We can say this model is the best since the p-value is below the significance level and with a high R^2 value showing almost all of the variance explained by the model.

# Question 16

# a. Logistic regression for hemorrhage  
glm\_hemorrhage <- glm(grmhem ~ apgar5 + tox, family = binomial, data = lowbwt)  
summary(glm\_hemorrhage)

##   
## Call:  
## glm(formula = grmhem ~ apgar5 + tox, family = binomial, data = lowbwt)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.2177 0.6268 -0.347 0.7283   
## apgar5 -0.2334 0.1051 -2.222 0.0263 \*  
## tox -1.2974 1.0778 -1.204 0.2287   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 84.542 on 99 degrees of freedom  
## Residual deviance: 76.967 on 97 degrees of freedom  
## AIC: 82.967  
##   
## Number of Fisher Scoring iterations: 5

# b. Predicted probability for apgar5 = 3 and tox = 1  
predict(glm\_hemorrhage, newdata = data.frame(apgar5 = 3, tox = 1), type = "response")

## 1   
## 0.09837279

# c. Hypothesis test for b1  
# Based on the model summary, we see b1 (agpar5) has a p-value below the significance level suggesting that the b1 differs significantly from zero. We reject the null hypothesis.

# Question 17

# Load dataset  
stenosis <- read\_excel("stenosis.xls")  
  
# a) Logistic regression model for aortic stenosis with smoking and gender  
model\_17a <- glm(disease ~ smoke + sex, family="binomial", data=stenosis)  
summary(model\_17a)

##   
## Call:  
## glm(formula = disease ~ smoke + sex, family = "binomial", data = stenosis)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.4882 0.2159 -2.261 0.0238 \*  
## smoke 0.1946 0.2903 0.670 0.5026   
## sex 0.7199 0.2881 2.499 0.0125 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 297.94 on 214 degrees of freedom  
## Residual deviance: 289.64 on 212 degrees of freedom  
## AIC: 295.64  
##   
## Number of Fisher Scoring iterations: 4

# Interpretation of odds  
exp(coef(model\_17a)[2]) # Odds ratio for smokers vs nonsmokers

## smoke   
## 1.214823

# b) Discuss interaction effects  
# Based on the model summary, I believe that the presence of aortic stenosis differs between males and females. The coefficient b2 for sex is below 0.05 suggesting statistical significance.

# Question 18

# Load dataset  
dialysis <- read\_excel("dialysis.xls")  
View(dialysis)  
  
# a) Logistic regression for age, sex, and race effects on peritonitis  
model\_18a\_age <- glm(perito ~ age, family="binomial", data=dialysis)  
model\_18a\_sex <- glm(perito ~ sex, family="binomial", data=dialysis)  
model\_18a\_race <- glm(perito ~ race, family="binomial", data=dialysis)  
summary(model\_18a\_age)

##   
## Call:  
## glm(formula = perito ~ age, family = "binomial", data = dialysis)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 0.770491 1.372803 0.561 0.575  
## age -0.002713 0.025542 -0.106 0.915  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 59.440 on 45 degrees of freedom  
## Residual deviance: 59.429 on 44 degrees of freedom  
## AIC: 63.429  
##   
## Number of Fisher Scoring iterations: 4

summary(model\_18a\_sex)

##   
## Call:  
## glm(formula = perito ~ sex, family = "binomial", data = dialysis)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 0.3102 0.3970 0.781 0.435  
## sex 0.7885 0.6513 1.211 0.226  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 59.440 on 45 degrees of freedom  
## Residual deviance: 57.919 on 44 degrees of freedom  
## AIC: 61.919  
##   
## Number of Fisher Scoring iterations: 4

summary(model\_18a\_race)

##   
## Call:  
## glm(formula = perito ~ race, family = "binomial", data = dialysis)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 0.4055 0.3450 1.175 0.240  
## race 1.0986 0.8545 1.286 0.199  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 59.440 on 45 degrees of freedom  
## Residual deviance: 57.542 on 44 degrees of freedom  
## AIC: 61.542  
##   
## Number of Fisher Scoring iterations: 4

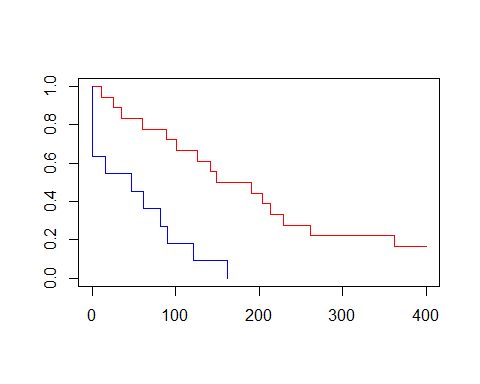
# b) Prediction for white male, age 50  
predict(model\_18a\_age, newdata=data.frame(age=50, sex=0, race=0), type="response")

## 1   
## 0.6535884

# c) Response Variable Categorization  
# Yes I do see a problem specifically with how age is categorized in comparison to the other variables. While the other variables are in binomial form (binary), the presence of a ordinal set of data is mixed with the bigger data meaning that it can obscure our model and some of our interpretations since it is not consistent.

# Question 19

# Load dataset  
cyto <- read\_excel("cyto.xls")  
view(cyto)  
  
# a) Survival curves for two treatment groups  
cyto\_surv <- Surv(cyto$time, cyto$censor)  
surv\_fit <- survfit(cyto\_surv ~ cyto$group)  
plot(surv\_fit, col=c("red", "blue"))



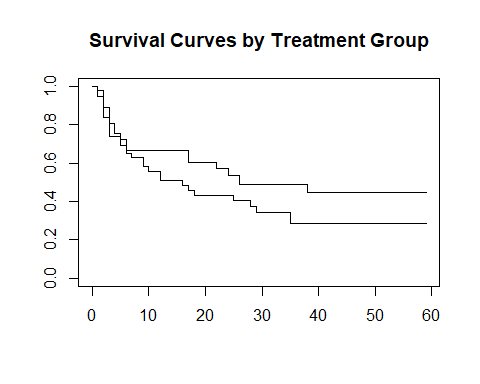
# b) Compare survival times visually  
# By inspecting the survival plot, we the blue treatment group has a much higher survival time almost 2x more than the other.   
  
# c) Log-rank test  
survdiff(cyto\_surv ~ cyto$group)

## Call:  
## survdiff(formula = cyto\_surv ~ cyto$group)  
##   
## N Observed Expected (O-E)^2/E (O-E)^2/V  
## cyto$group=1 18 15 21.37 1.90 12.4  
## cyto$group=2 11 11 4.63 8.74 12.4  
##   
## Chisq= 12.4 on 1 degrees of freedom, p= 4e-04

# Based on the summary, we see a p-value below the significance value therefore we reject the null hypothesis.

# Question 20

# a) Survival curves by treatment group  
bladder\_surv <- survfit(Surv(time, censor) ~ group, data = bladder)  
plot(bladder\_surv, main = "Survival Curves by Treatment Group")



# b) Based on the plot, it doesn't seem like the individuals from one group have a significantly longer time to first recurrence of tumor. As we can tell, the behaviors and trends of both groups are very similar.  
  
# c)   
survdiff(Surv(time, censor) ~ group, data = bladder)

## Call:  
## survdiff(formula = Surv(time, censor) ~ group, data = bladder)  
##   
## N Observed Expected (O-E)^2/E (O-E)^2/V  
## group=1 48 29 24.9 0.671 1.52  
## group=2 38 18 22.1 0.757 1.52  
##   
## Chisq= 1.5 on 1 degrees of freedom, p= 0.2

# Based on the summary p-value, it is above the significance level and we fail to reject the null hypothesis that the recurrence times are identical between the two groups.