PHP2514_Basso_HW2_2021

October 26, 2021

1 PHP2514: Applied Generalized Linear Models

1.1 Homework 2

Antonella Basso

1.1.1 Question 1:

The dataset "Optics.csv" contains information from a math education graduate student research project. For the optics module in a high school freshman physical science class, the randomized study compared two instruction methods (1=model building inquiry, 0=traditional scientific). The response variable was an optics post-test score ("OptPost"). Other explanatory variables were gender (1=female, 0=male) and the optics pre-test score ("OptPre"). The primary research question was to test the effectiveness of the new method (model building inquiry ("MBI")) over the traditional one.

a) Conduct a comprehensive Exploratory Data Analysis (EDA) to inspect, understand and describe the information collected in this dataset. Use appropriate summary statistics and plots to present your results from the EDA.

b) What do you think about the effectiveness of the two instruction methods based on the results of the EDA?

c) Choose a model selection procedure (backward, forward, or stepwise) to find the model the best fits your data. Check for all possible interactions terms among the covariates in the model.

d) Based on the conclusions from the model selection procedure and your personal judgement (e.g., adjusting for any important (according to your opinion) covariates), state the form of the model that best describes your data. What are the assumptions of this model?

e) Assess the overall fit of the model using regression diagnostics to check model assumptions, identify questionable observations (outliers, influential points), and/or find other problems indicating model inadequacy.

f) Suggest a way to assess the predictive accuracy of your "best" model. Clearly describe your approach, conduct the necessary analysis and comment on the results.

g) Interpret the regression coefficients of your "best" model.

h) Based on the analysis results what is your overall conclusion regarding the primary research question of this study?

[1]:
```
#DATA WRANGLING

#installing tidyverse
install.packages("tidyverse")
library(tidyverse)
```

```
Updating HTML index of packages in '.Library'

Making 'packages.html' …
 done

Warning message in system("timedatectl", intern = TRUE):
"running command 'timedatectl' had status 1"
  Attaching packages                           tidyverse
1.3.1

 ggplot2 3.3.5      purrr   0.3.4
 tibble  3.1.4      dplyr   1.0.7
 tidyr   1.1.3      stringr 1.4.0
 readr   1.4.0      forcats 0.5.1

  Conflicts
tidyverse_conflicts()
 dplyr::filter() masks stats::filter()
 dplyr::lag()    masks stats::lag()
```

[2]:
```
#importing "Optics" data
df <- read.csv("/home/jovyan/AGLM/HW2/Optics.csv")
head(df)
```

A data.frame: 6 × 5

	ID <int>	OptPost <int>	method <chr>	gender <int>	OptPre <int>
1	4	50	MBI	0	50
2	5	67	MBI	0	50
3	6	61	MBI	0	30
4	8	92	MBI	0	67
5	12	59	MBI	1	42
6	13	16	MBI	1	8

[3]:
```
#removing ID column for simplicity
optics <- subset(df, select = -ID)

#renaming method values (MBI:0, traditional:1)
#this adds a column that gives a 1 if the group is traditional and 0 otherwise
#NOTE: 1*TRUE=1, 1*FALSE=0
optics$method_bin <- as.factor(1*(optics$method=='traditional'))
head(optics)
```

A data.frame: 6 × 5

	OptPost <int>	method <chr>	gender <int>	OptPre <int>	method_bin <fct>
1	50	MBI	0	50	0
2	67	MBI	0	50	0
3	61	MBI	0	30	0
4	92	MBI	0	67	0
5	59	MBI	1	42	0
6	16	MBI	1	8	0

a) Exploratory Data Analysis (EDA) The variables in this dataset are as follows: - Outcome (Y: OptPost score) - Covariate (X_1: OptPre score, X_2: method (0=MBI, 1=traditional), X_3: gender (0=male, 1=female))

The primary outcome of interest is a discrete random variable with ordinal scale, while the predictor variables (covariates) are discrete with ordinal scale (X_1); and categorical with nominal scale or binary (X_2 and X_3).

This EDA consists of: - Descriptive Statistics - Boxplots - Histograms - Q-Q Plot - Scatter Plot

optics

Descriptive Statistics:

- Summary of OptPost (minimum value, 1st quartile, median, mean, 3rd quartile, maximum value) for the whole data and for each method group
- Standard deviation (SD) and variance of OptPost data for each method group

[4]:
```
#summary of total OptPost and OptPost by group (MBI and traditional)
summary(optics$OptPost)
by(optics$OptPost, optics$method, summary, na.rm=TRUE)

#SD and variance of OptPost for MBI and traditional groups
sd(optics[optics$method == "MBI",]$OptPost)
var(optics[optics$method == "MBI",]$OptPost)
sd(optics[optics$method == "traditional",]$OptPost)
var(optics[optics$method == "traditional",]$OptPost)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  16.00   46.00   59.00   58.78   69.00   92.00 
optics$method: MBI
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  16.00   52.00   61.00   63.43   84.00   92.00 
------------------------------------------------------------
optics$method: traditional
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  20.00   45.50   50.00   52.69   58.00   84.00 
```

19.0776608329518

363.957142857143

15.4000811686173

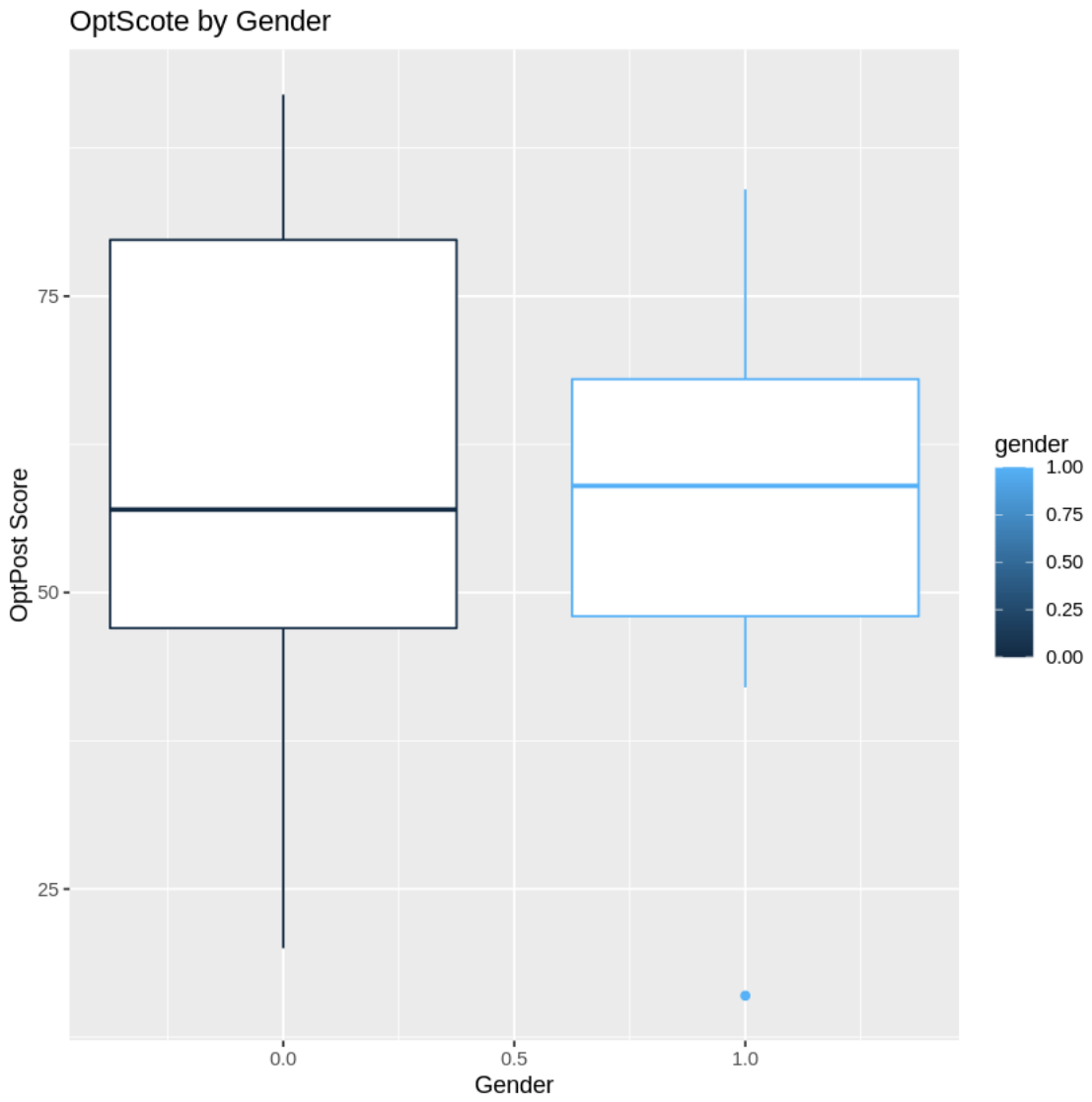
237.1625

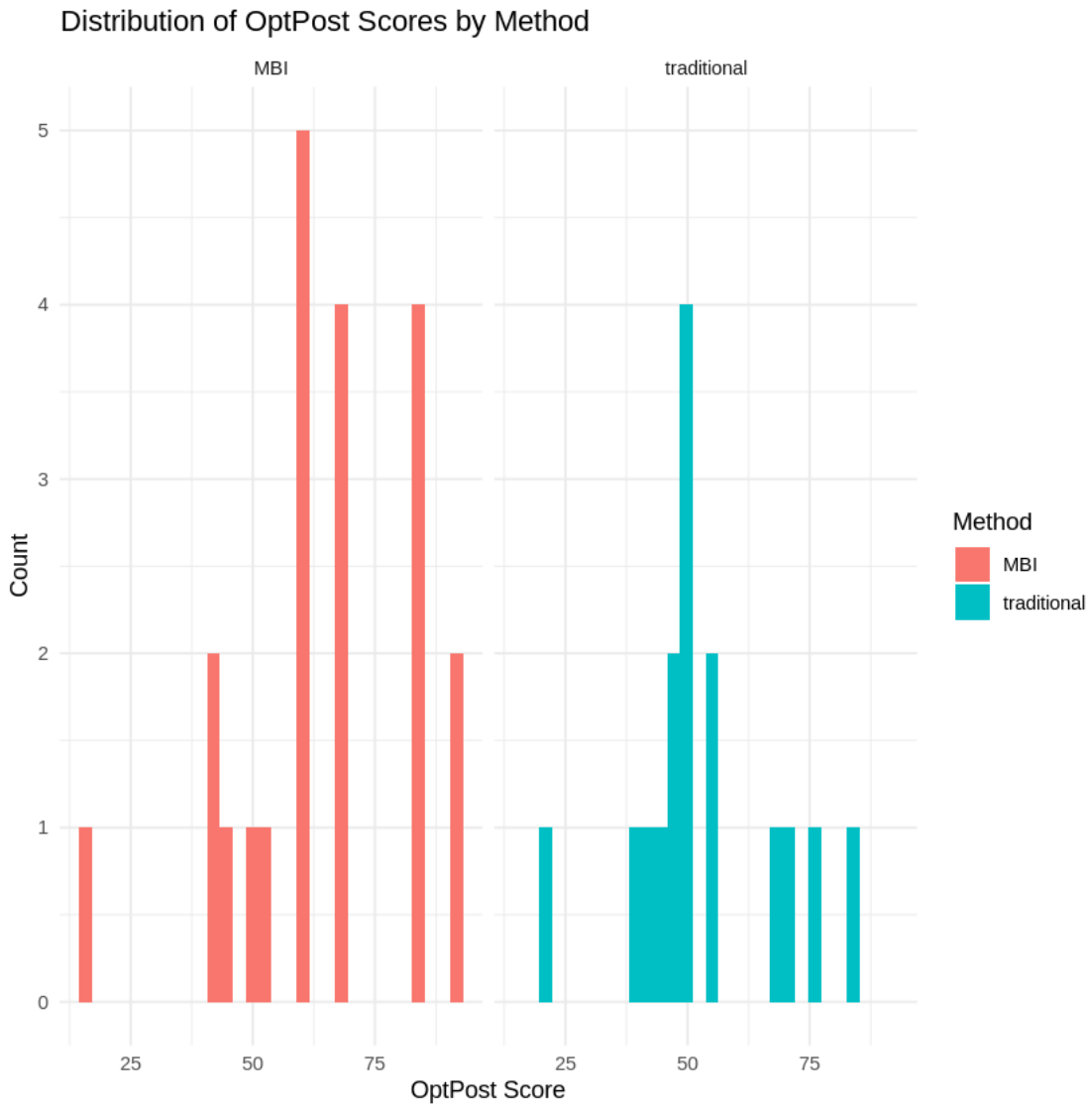
Graphs/Plots:

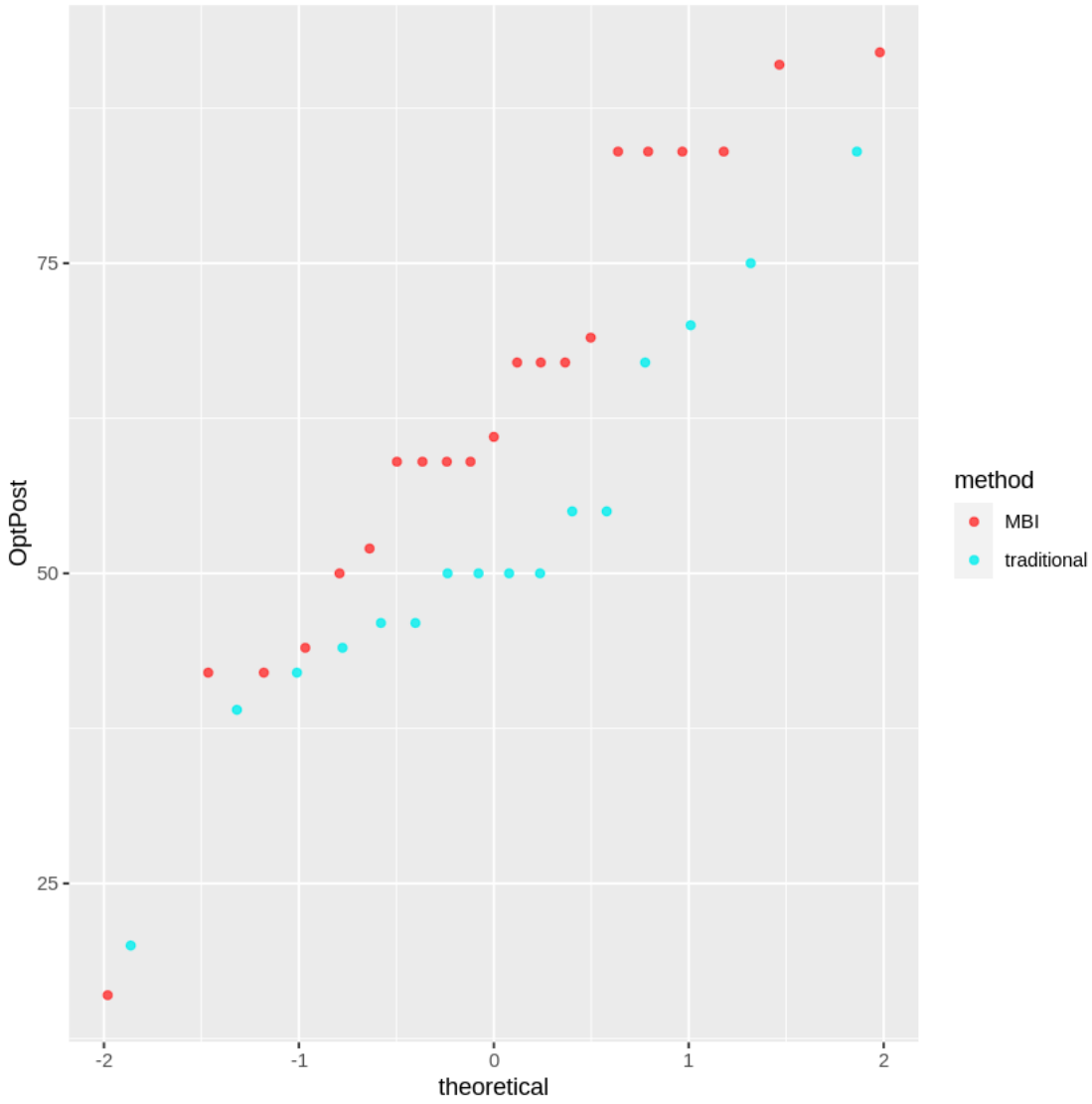
- Boxplots (show a side by side comparison of the mean and spread of OptPost score data for males (0) and females (1))
- Histograms (show a side by side comparison of the distribution (appearing normal) of OptPost scores for each method group)
- Q-Q Plot (shows that the response variable (OptPost score) is approximately (not perfectly) normally distributed and that the distribution for each method group is roughly the same (this is can be observed by noticing the parallel-like linear trend in the plot) despite the values not being the same)
- Scatter Plot (shows the relationship between OptPre and OptPost scores)

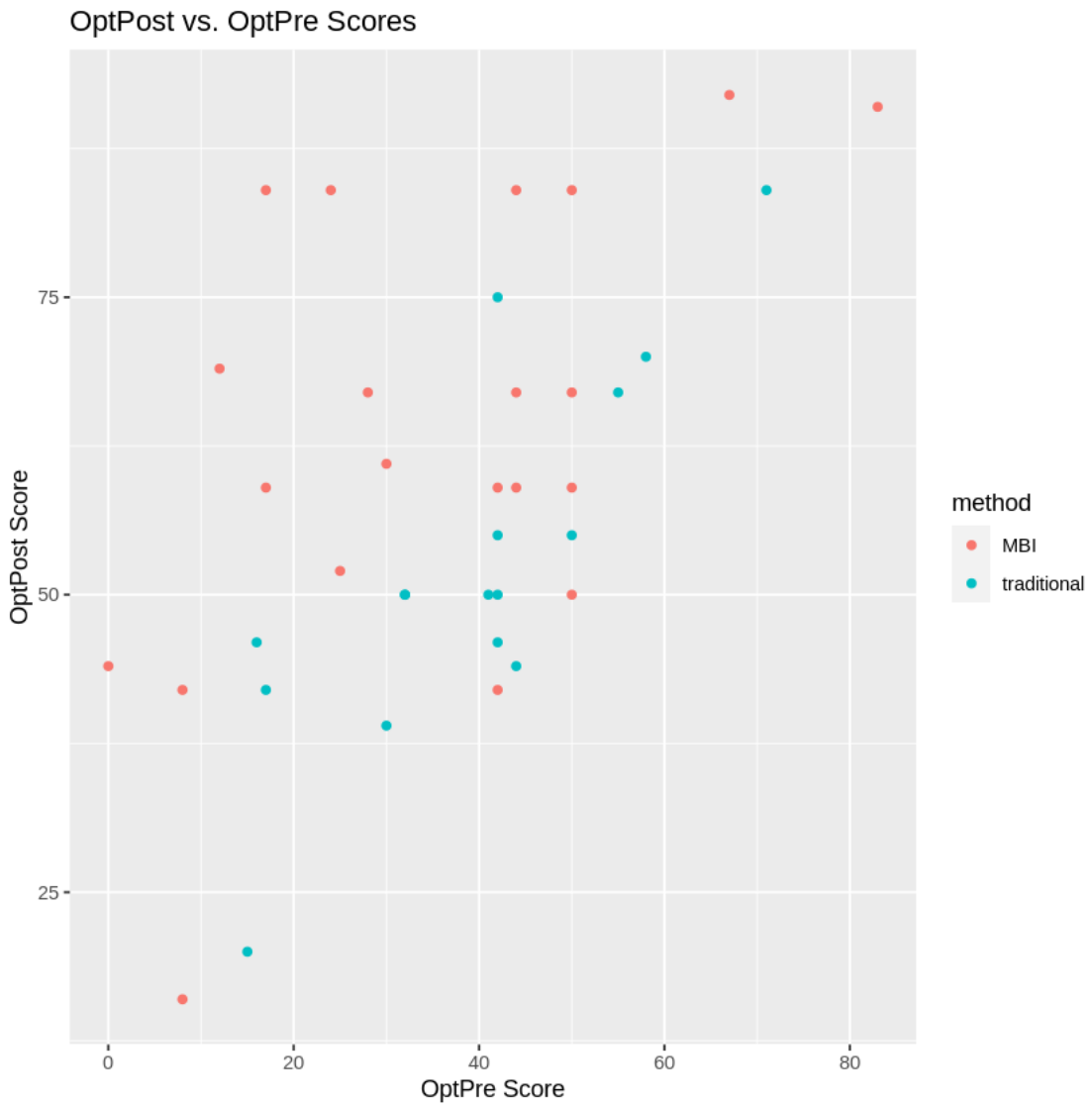
[5]:
```
#Boxplots
ggplot(optics, aes(group=gender, x=gender, y=OptPost, color=gender)) +␣
↪geom_boxplot() +
labs(x = "Gender", y = "OptPost Score", title = "OptScote by Gender")

#Histograms
ggplot(data=optics, aes(x=OptPost, fill=method)) +
  geom_histogram() +
  scale_fill_discrete(name = "Method") +
  labs(x="OptPost Score", y = "Count", title = "Distribution of OptPost Scores␣
↪by Method") +
  facet_wrap(~method) +
  theme_minimal()

#Q-Q Plot
ggplot(optics, aes(sample = OptPost)) + stat_qq(aes(color = method), alpha = 0.
↪8) + scale_color_manual(values =c("firebrick1", "cyan2")) + labs(y =␣
↪"OptPost")

#Scatterplot
ggplot(optics) + geom_point(aes(x = OptPre, y = OptPost, color=method)) +
labs(x = "OptPre Score", y = "OptPost Score", title = "OptPost vs. OptPre␣
↪Scores")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


OptScote by Gender
OptPost Score
75
50
25
0.0
0.5
1.0
Gender
gender
1.00
0.75
0.50
0.25
0.00


Distribution of OptPost Scores by Method
MBI
traditional
Count
OptPost Score
0
1
2
3
4
5
25
50
75
Method
MBI
traditional


OptPost
75
50
25
-2
-1
0
1
2
theoretical
method
MBI
traditional



b) Interpretation of Results: Based on this EDA, it appears that those that recieve the MBI instruction method, on average, tend to produce slightly higher OptPost scores than those who recieve the traditional instruction method. Moreover, test sores seem to be normally distributed with a potential right skew (given the restults from the Q-Q plot above and the fact that sample means are greater than the medians). Additionally, there seems to be a relationship (more linear for the traditional group than the MBI group) between OptPre and OptPost scores. Specifically, it appears that those who perform well on the OptPre tend to also perform well on the OptPost. Given the relative position and spread of the data however, it is possible for some who perform poorly on the OptPre to perform well on the OptPost (but, this is more so the case for those in the MBI group, as can be observed in the scatter plot above). Lastly, as can be seen on the boxplots above, there seems to be little to no relationship between gender and OptPost performance.

c) Model Selection:

- Backward Selection

[6]:
```
#Method 1: FWD, BWD, SW regression (http://www.sthda.com/english/articles/
↪37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/)
#library(MASS)

#opt_model <- lm(OptPost ~ OptPre + gender + method_bin, data = optics)

#fit1 <- lm(OptPost ~ ., optics) #full model
#fit2 <- lm(OptPost ~ 1, optics) #null model
#stepAIC(fit1,direction="backward")
#stepAIC(fit2,direction="forward",scope=list(upper=fit1,lower=fit2))
#stepAIC(fit2,direction="both",scope=list(upper=fit1,lower=fit2))


#Method 2: FWD, BWD, SW regression (#https://stackoverflow.com/questions/
↪55821462/
↪how-can-i-perform-a-forward-selection-backward-selection-and-stepwise-regressi)
#Full and null models
#nullmod <- lm(OptPost ~ 1, data = optics)
#fullmod <- lm(OptPost ~ ., data = optics)

#Forward
#reg1A <- step(nullmod, scope = list(lower = nullmod, upper = fullmod),␣
↪direction="forward")
#reg1A
#summary(reg1A)

#Backward
#reg1B <- step(nullmod, scope = list(lower = fullmod, upper = nullmod),␣
↪direction="backward")
#reg1B
#summary(reg1B)

#Stepwise
#reg1C <- step(nullmod, scope = list(lower = fullmod, upper = nullmod),␣
↪direction="both")
#reg1C
#summary(reg1C)
```

[7]:
```
#Method 3: BWD selection (in-class)

#saturated model (all possible interaction terms)
m1 <- glm(OptPost ~ OptPre*method*gender, family=gaussian, data=optics)
summary(m1)

#model 2 (only 2-way interaction terms included)
```

```r
m2 <- glm(OptPost ~ OptPre*method + OptPre*gender + method*gender,␣
 ↪family=gaussian, data=optics)
summary(m2)

#model 3 (removing the least beneficial 2-way interaction term)
m3 <- glm(OptPost ~ method*OptPre + method*gender, family=gaussian, data=optics)
summary(m3)

#model 4 (removing the least beneficial 2-way interaction term)
m4 <- glm(OptPost ~ method*gender + OptPre, family=gaussian, data=optics)
summary(m4)

#model 5 (removing gender covariate, and hence, the last interaction term -␣
 ↪only OptPre and method covariates left)
m5 <- glm(OptPost ~ OptPre + method, family=gaussian, data=optics)
summary(m5)
```

```
Call:
glm(formula = OptPost ~ OptPre * method * gender, family = gaussian,
    data = optics)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-26.7906   -6.6161   -0.9723    9.0781   24.8978

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                     51.6253     7.1873   7.183  6.6e-08 ***
OptPre                           0.4398     0.1643   2.676  0.01211 *
methodtraditional              -42.5676    14.9409  -2.849  0.00798 **
gender                         -13.3555    13.4720  -0.991  0.32971
OptPre:methodtraditional         0.5559     0.3499   1.589  0.12297
OptPre:gender                    0.1253     0.3680   0.340  0.73596
methodtraditional:gender        38.1092    23.3951   1.629  0.11414
OptPre:methodtraditional:gender -0.4999     0.5855  -0.854  0.40020
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 184.936)

    Null deviance: 11884.3  on 36  degrees of freedom
Residual deviance:  5363.1  on 29  degrees of freedom
AIC: 307.13

Number of Fisher Scoring iterations: 2
```

```
Call:
glm(formula = OptPost ~ OptPre * method + OptPre * gender + method *
    gender, family = gaussian, data = optics)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-30.486   -6.159   -1.405    7.996   25.695

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               50.15922    6.94762   7.220 4.91e-08 ***
OptPre                     0.47919    0.15701   3.052  0.00473 **
methodtraditional        -35.56522   12.43227  -2.861  0.00763 **
gender                    -6.92954   11.12309  -0.623  0.53800
OptPre:methodtraditional   0.37738    0.27930   1.351  0.18674
OptPre:gender             -0.07217    0.28494  -0.253  0.80177
methodtraditional:gender  19.82508    9.37783   2.114  0.04293 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 183.2654)

    Null deviance: 11884  on 36  degrees of freedom
Residual deviance:  5498  on 30  degrees of freedom
AIC: 306.05

Number of Fisher Scoring iterations: 2



Call:
glm(formula = OptPost ~ method * OptPre + method * gender, family = gaussian,
    data = optics)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-29.135   -6.695   -1.938    7.737   25.403

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               50.6951     6.5170   7.779  8.9e-09 ***
methodtraditional        -34.7801    11.8565  -2.933  0.00626 **
OptPre                     0.4648     0.1441   3.225  0.00297 **
gender                    -9.2782     6.0499  -1.534  0.13527
methodtraditional:OptPre   0.3586     0.2651   1.352  0.18604
methodtraditional:gender  19.3443     9.0441   2.139  0.04043 *
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 177.7329)

    Null deviance: 11884.3  on 36  degrees of freedom
Residual deviance:  5509.7  on 31  degrees of freedom
AIC: 304.13

Number of Fisher Scoring iterations: 2



Call:
glm(formula = OptPost ~ method * gender + OptPre, family = gaussian,
    data = optics)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-26.658   -7.077   -2.749    5.773   27.547

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              46.7495     5.9025   7.920 4.87e-09 ***
methodtraditional       -20.7871     5.8637  -3.545  0.00123 **
gender                   -8.6577     6.1101  -1.417  0.16616
OptPre                    0.5708     0.1225   4.658 5.36e-05 ***
methodtraditional:gender 18.4552     9.1362   2.020  0.05182 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 182.3366)

    Null deviance: 11884.3  on 36  degrees of freedom
Residual deviance:  5834.8  on 32  degrees of freedom
AIC: 304.25

Number of Fisher Scoring iterations: 2



Call:
glm(formula = OptPost ~ OptPre + method, family = gaussian, data = optics)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-31.5572   -8.2133    0.5106    7.0163   31.1524

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)        42.8545     5.3352   8.032 2.31e-09 ***
OptPre              0.5878     0.1254   4.690 4.32e-05 ***
methodtraditional -13.2761     4.6481  -2.856  0.00726 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 193.54)

    Null deviance: 11884.3  on 36  degrees of freedom
Residual deviance:  6580.4  on 34  degrees of freedom
AIC: 304.7

Number of Fisher Scoring iterations: 2
```

d) Model Selection Conclusion: Based on the backward selection procedure, models 3, 4, and 5 have the smallest AIC values with AIC's of 304.13, 304.25, and 304.7 respectively. We can assume that the difference in these AIC values comes from removing terms. However, given that this difference is small, the intuition (from the EDA and model summaries) that gender plays little if any role in OptPost scores (yet is included in models 3 and 4), and the fact that model 5 is the most parsimonious of all models, it is safe to assume that model 5 fits and describes the data best:

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 = 42.8545 + 0.5878 X_1 - 13.2761 X_2.$$

e) Checking Model Fit & Model Assumptions:

Model Fit: Assessing the overall fit of the model: - GoF/Deviance: Comparing chosen model to the saturated model - Wald Test: Checking individual parameters - ANOVA: Comparing linear model to the saturated linear model (F-test) - R, R^2, Adjusted R^2 - AIC/BIC

Model Assumptions:

- **Linearity** (functional form for linear models): If Y and X have a linear relationship, there should be no significant trend in the residuals with respect to Y. This can be assessed by checking if a flexible fit of the mean of the residuals is constant (usually semi-straight line around the mean).
- **Normality**: If residuals are normally distributed, the respective QQ-plot will display values roughly along the diagonal line, especially near the center. The Shapiro–Wilk test is a formal way to test for normality of residuals (we must have a p-value greater than 0.05, since the null hypothesis assumes normality).
- **Homoscedasticity** (Constant Variance): If residuals are homoscedastic, we expect a flexible fit of the mean of the transformed residuals (red line) to be almost constant about 1. But, if there are clear non-constant patterns, then there is evidence of heteroscedasticity. A formal test to check the null hypothesis of homoscedasticity in the residuals is the Breusch–Pagan test (we must have a p-value greater than 0.05, since the null hypothesis assumes homoscedasticity, that is, that the error variances are all equal).

- **Multicollinearity**: If continuous predictors are not linearly related, then VIF is close to 1 (less than 5). To test for multicollinearity between categorical variables we use a Chi-Square test for independence. And, to test for multicollinearity between categorical and continuous variables, we use ANOVA or a t-test (depending on the number of groups in the categorical variable).
- **Outliers and Influential Points**: If there are no outliers (observations with a response far away from the regression plane), then the standardized residual of an observation is less than 3 in absolute value. If there are no influential/high-leverage points (observations with a relatively large effect on estimates of model coefficients), then the Cook's distance is less than 1. Both outliers and high-leverage points can be identified with the residuals vs. leverage plot.

Further:

- Functional forms of model covariates: Linearity (above)
- Adequate fit of the covariates in the model: Comparing standardized residuals with covariates IN and NOT IN the model (to check for linearity and possible missing information)

Reference: https://bookdown.org/egarpor/PM-UC3M/lm-ii-diagnostics.html

Model Fit: Assess the overall fit of the model: - GoF/Deviance: observations - Wald Test: observations - ANOVA: observations - R, R^2, Adjusted R^2: observations - AIC/BIC: observations

[8]:
```
#GLM deviances, Wald tests, AIC, ANOVA

summary(m5) #chosen model
summary(m1) #saturated model

anova(m5, m1, test="F") #f test (more appropriate for this glm - gaussian)
anova(m5, m1, test="LRT") #log-likelihood ratio test
```

```
Call:
glm(formula = OptPost ~ OptPre + method, family = gaussian, data = optics)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-31.5572   -8.2133    0.5106    7.0163   31.1524

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        42.8545     5.3352   8.032 2.31e-09 ***
OptPre              0.5878     0.1254   4.690 4.32e-05 ***
methodtraditional -13.2761     4.6481  -2.856  0.00726 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 193.54)
```

```
    Null deviance: 11884.3  on 36  degrees of freedom
Residual deviance:  6580.4  on 34  degrees of freedom
AIC: 304.7

Number of Fisher Scoring iterations: 2
```

```
Call:
glm(formula = OptPost ~ OptPre * method * gender, family = gaussian,
    data = optics)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-26.7906   -6.6161   -0.9723    9.0781   24.8978

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                     51.6253     7.1873   7.183  6.6e-08 ***
OptPre                           0.4398     0.1643   2.676  0.01211 *
methodtraditional              -42.5676    14.9409  -2.849  0.00798 **
gender                         -13.3555    13.4720  -0.991  0.32971
OptPre:methodtraditional         0.5559     0.3499   1.589  0.12297
OptPre:gender                    0.1253     0.3680   0.340  0.73596
methodtraditional:gender        38.1092    23.3951   1.629  0.11414
OptPre:methodtraditional:gender -0.4999     0.5855  -0.854  0.40020
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 184.936)

    Null deviance: 11884.3  on 36  degrees of freedom
Residual deviance:  5363.1  on 29  degrees of freedom
AIC: 307.13

Number of Fisher Scoring iterations: 2
```

A anova: 2 × 6

	Resid. Df <dbl>	Resid. Dev <dbl>	Df <dbl>	Deviance <dbl>	F <dbl>	Pr(>F) <dbl>
1	34	6580.359	NA	NA	NA	NA
2	29	5363.143	5	1217.216	1.316365	0.2847214

A anova: 2 × 5

	Resid. Df <dbl>	Resid. Dev <dbl>	Df <dbl>	Deviance <dbl>	Pr(>Chi) <dbl>
1	34	6580.359	NA	NA	NA
2	29	5363.143	5	1217.216	0.2536434

[9]:
```r
#Linear model ANOVA, R, R^2, adjusted R^2

lm5 <- lm(OptPost ~ OptPre + method, data=optics) #chosen model
summary(lm5)
lm1 <- lm(OptPost ~ OptPre*gender*method, data=optics) #saturated model
summary(lm1)

anova(lm5, lm1, test="F")
anova(lm5, lm1, test="LRT")
```

```
Call:
lm(formula = OptPost ~ OptPre + method, data = optics)

Residuals:
     Min       1Q   Median       3Q      Max
-31.5572  -8.2133   0.5106   7.0163  31.1524

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        42.8545     5.3352   8.032 2.31e-09 ***
OptPre              0.5878     0.1254   4.690 4.32e-05 ***
methodtraditional -13.2761     4.6481  -2.856  0.00726 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.91 on 34 degrees of freedom
Multiple R-squared:  0.4463,        Adjusted R-squared:  0.4137
F-statistic:  13.7 on 2 and 34 DF,  p-value: 4.322e-05



Call:
lm(formula = OptPost ~ OptPre * gender * method, data = optics)

Residuals:
     Min       1Q   Median       3Q      Max
-26.7906  -6.6161  -0.9723   9.0781  24.8978

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    51.6253     7.1873   7.183  6.6e-08 ***
OptPre                          0.4398     0.1643   2.676  0.01211 *
gender                        -13.3555    13.4720  -0.991  0.32971
methodtraditional             -42.5676    14.9409  -2.849  0.00798 **
OptPre:gender                   0.1253     0.3680   0.340  0.73596
OptPre:methodtraditional        0.5559     0.3499   1.589  0.12297
gender:methodtraditional       38.1092    23.3951   1.629  0.11414
```

```
OptPre:gender:methodtraditional  -0.4999     0.5855  -0.854  0.40020
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.6 on 29 degrees of freedom
Multiple R-squared:  0.5487,        Adjusted R-squared:  0.4398
F-statistic: 5.037 on 7 and 29 DF,  p-value: 0.0007952
```

A anova: 2 × 6

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	34	6580.359	NA	NA	NA	NA
2	29	5363.143	5	1217.216	1.316365	0.2847214

A anova: 2 × 5

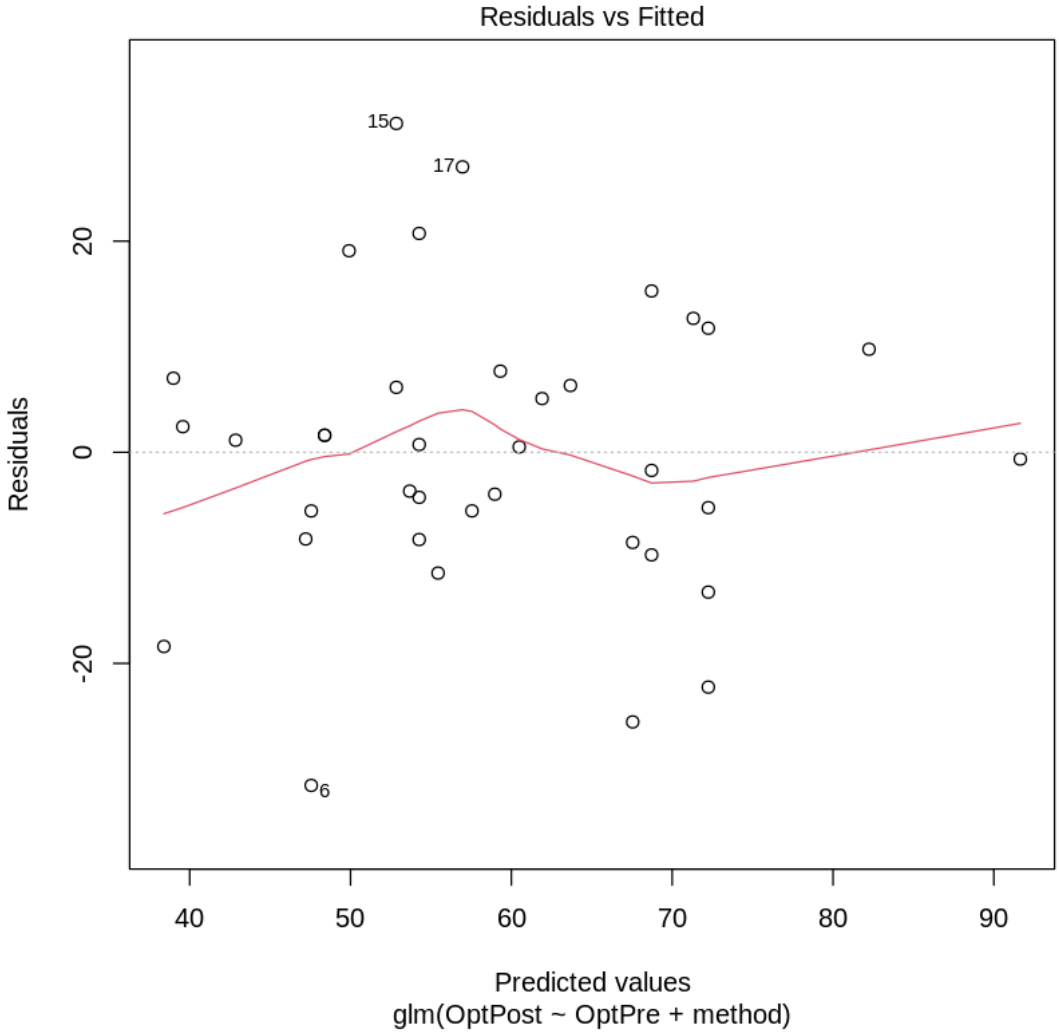
	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	Pr(>Chi) <dbl>
1	34	6580.359	NA	NA	NA
2	29	5363.143	5	1217.216	0.2536434

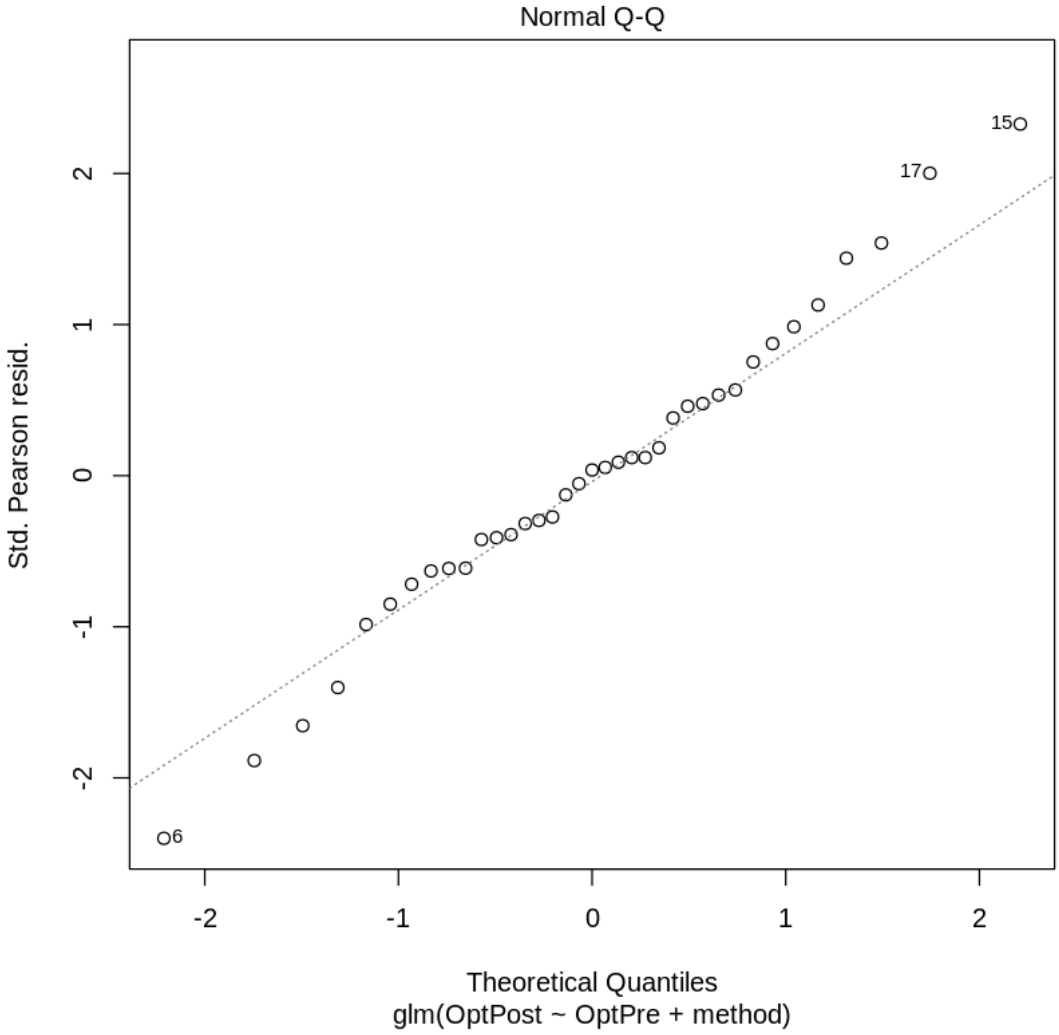
Model Assumptions:

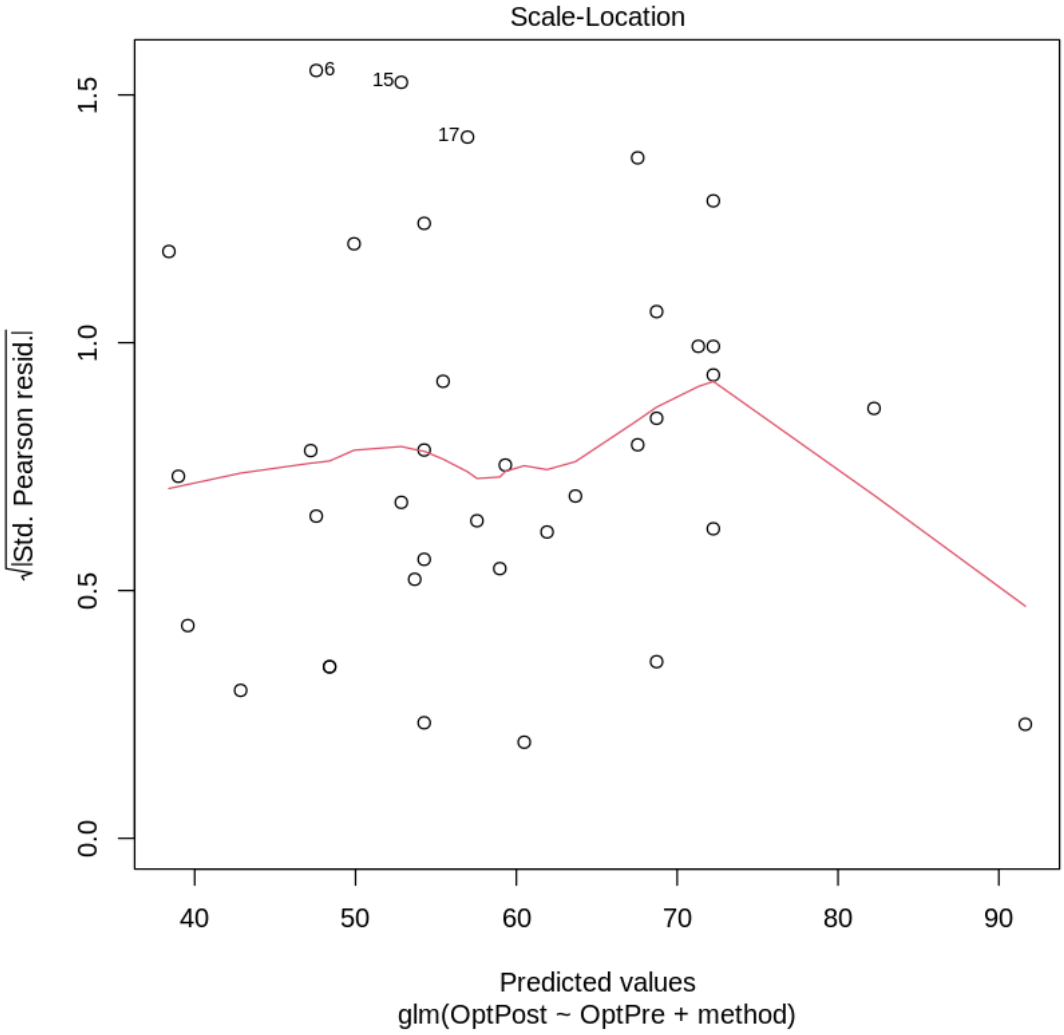
- **Linearity**: There is no evident linearity between the fitted values and residuals (plot 1). Thus, we may conclude that this assumption holds.
- **Normality**: The Q-Q plot (plot 2) and Shapiro-Wilk normality test demonstrate that residuals are normally distributed. Thus, we may conclude that this assumption holds.
- **Homoscedasticity** (Constant Variance): The abscence of a semi-straight line near 1 (plot 3), tells us that we might have some heteroscedasticity. However, the Breusch–Pagan test with a p-value of 0.068 (greater than 0.05) confirms that we may not reject the null hypothesis of homoscedasticity, and hence this assumption holds.
- **Multicollinearity**: The t-test to observe group differences between methods in terms of OptPre scores, demonstrates that there is no significant difference between groups (p-value greater than 0.05), and hence, the two variables are not correlated/collinear. Thus, we may conclude that this assumption holds.
- **Outliers and Influential Points**: Given that the absolute value of standardized residuals do not exceed 3 (plot 4) and the Cook's distance is less than 1 (plot 5), we may conclude that there are no outliers or influential points. Thus, this assumption holds.

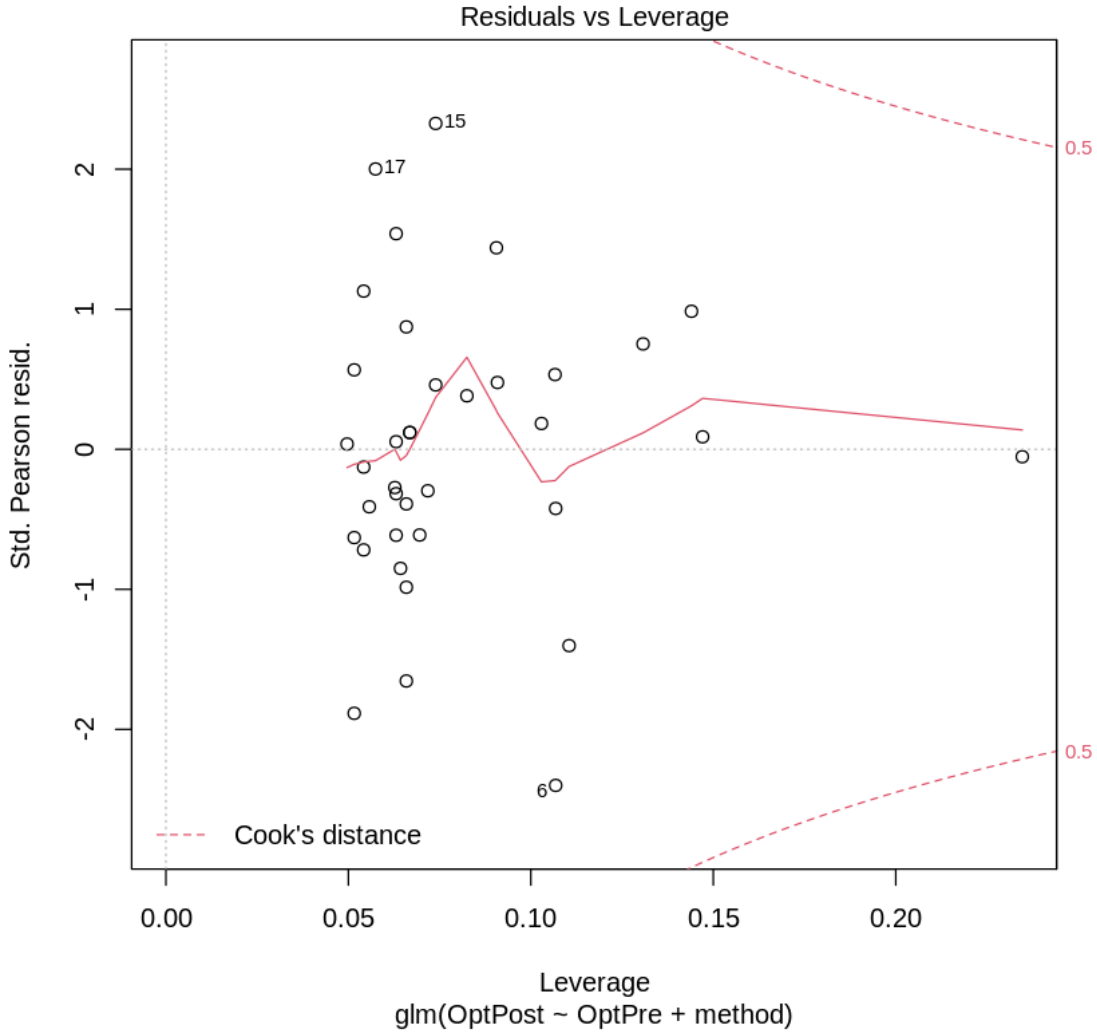
[10]:
```
#CHECKING MODEL ASSUMPTIONS

#Linearity (plot 1)
#no linearity between fitted values and residuals
#therefore, assumption is correct

#Normality (plot 2 and Shapiro-Wilk test)
#residuals are normally distributed in the QQ-Plot
#therefore, assumption is correct

#Homoscedasticity (plot 3 and Breusch-Pagan test)
#not really a semi-straight line near 1
```

```r
#p>0.05
#therefore, assumption is correct

#Multicollinearity (t-test)
#between 2 predictor variables (1 categorical and 1 numerical)
#p>0.05, there is no significant difference between groups and hence, they are␣
 ↪not correlated/collinear
#therefore, assumption is correct

#Outliers and Influential Points (plots 4 and 5)
#|standardized residuals| not greater than 3 and cook's distance is less than 1
#no outliers or influential points
#therefore, assumption is correct


plot(m5)
```


Residuals vs Fitted
Residuals
20
0
-20
15
17
6
40
50
60
70
80
90
Predicted values
glm(OptPost ~ OptPre + method)


Normal Q-Q
Std. Pearson resid.
Theoretical Quantiles
glm(OptPost ~ OptPre + method)
15
17
6
-2
-1
0
1
2


Scale-Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
6
15
17
40
50
60
70
80
90
Predicted values
glm(OptPost ~ OptPre + method)



[11]:
```
#Normality
#Null: data comes from a normally distributed population
#p>0.05
#therefore, assumption is correct

shapiro.test(rstandard(m5))
```

```
        Shapiro-Wilk normality test

data:  rstandard(m5)
W = 0.98699, p-value = 0.9355
```

[12]:
```r
#Homoscedasticity
#Null: assumes error variances are all equal (homoscedasticity)
#p>0.05
#therefore, assumption is correct

library(lmtest)
bptest(m5)
```

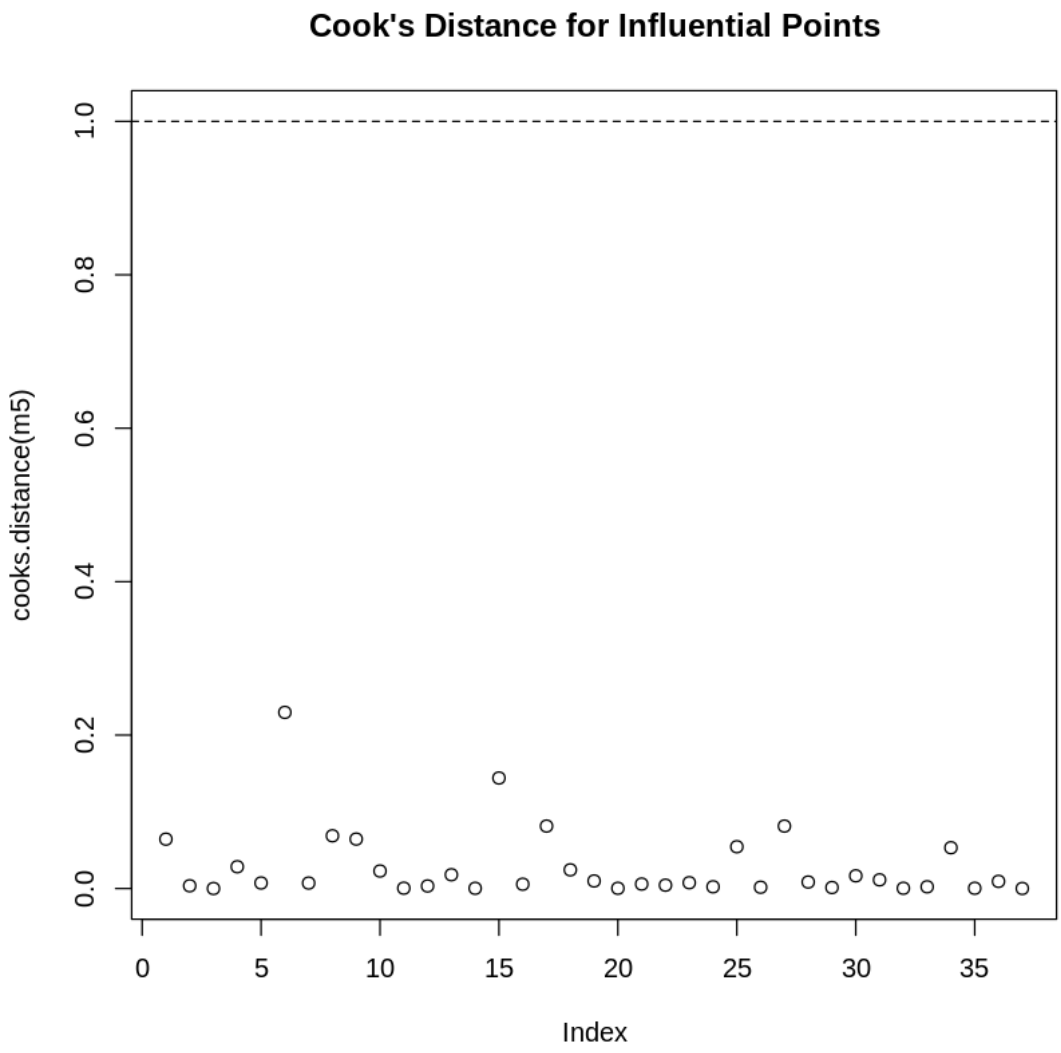
```
Loading required package: zoo


Attaching package: ‘zoo’


The following objects are masked from ‘package:base’:

    as.Date, as.Date.numeric



        studentized Breusch-Pagan test

data:  m5
BP = 5.3782, df = 2, p-value = 0.06794
```

[13]:
```r
#Multicollinearity
#between 2 predictor variables (1 categorical and 1 numerical)
#p>0.05, there is no significant difference between groups and hence, they are␣
 ↪not correlated/collinear
#therefore, assumption is correct

mbi_df <- optics[which(optics$method == "MBI"), ]
t_df <- optics[which(optics$method == "traditional"), ]

t.test(mbi_df$OptPre, t_df$OptPre) #OptPre comparison for instruction methods
```

```
        Welch Two Sample t-test

data:  mbi_df$OptPre and t_df$OptPre
t = -0.72132, df = 34.987, p-value = 0.4755
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -16.449942   7.824942
sample estimates:
mean of x mean of y
```

```
   35.0000    39.3125
```

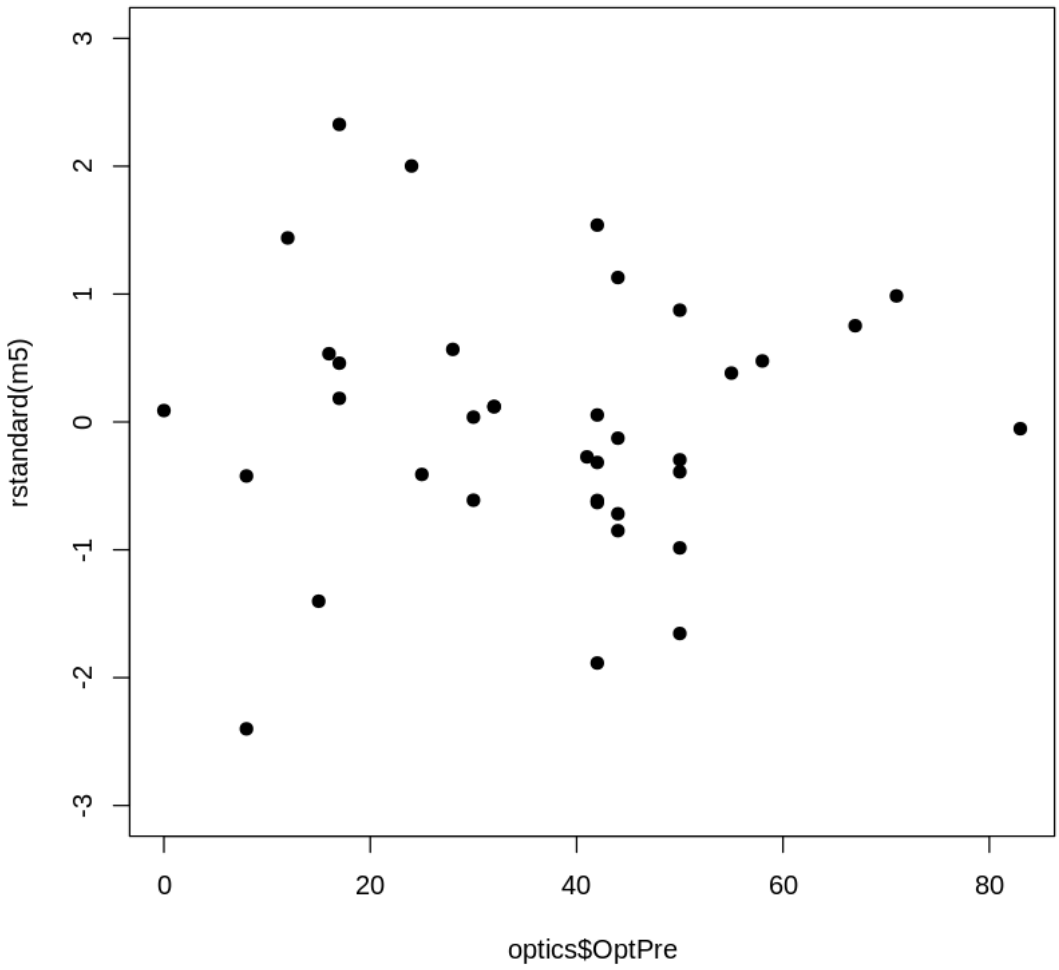
[14]:
```
#Influential points
#no values exceed the threshold (1)
#therefore, assumption is correct

plot(cooks.distance(m5), ylim=c(0,1), main = "Cook's Distance for Influential␣
 ↪Points")
abline(h = 1, lty = 2) #cutoff line at 1 (degress of freedom/number of␣
 ↪observations is close to 1)
```

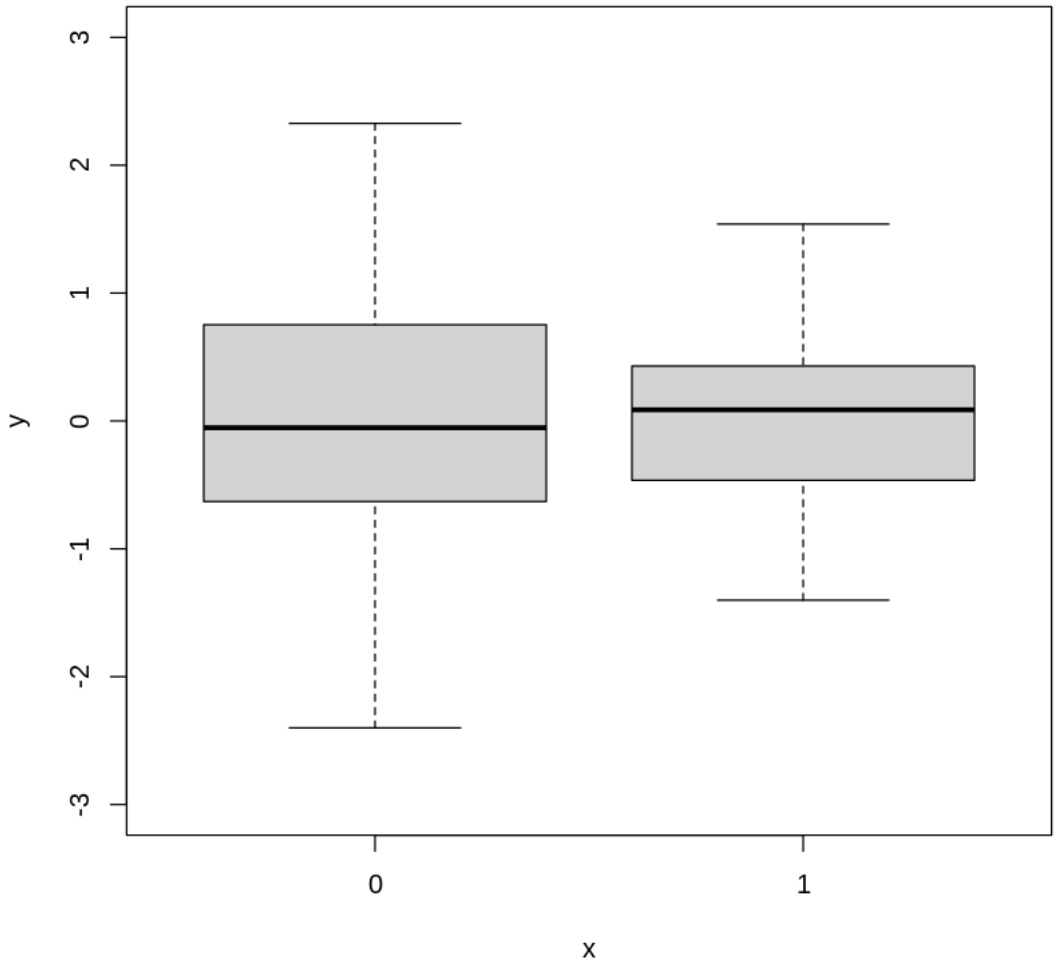


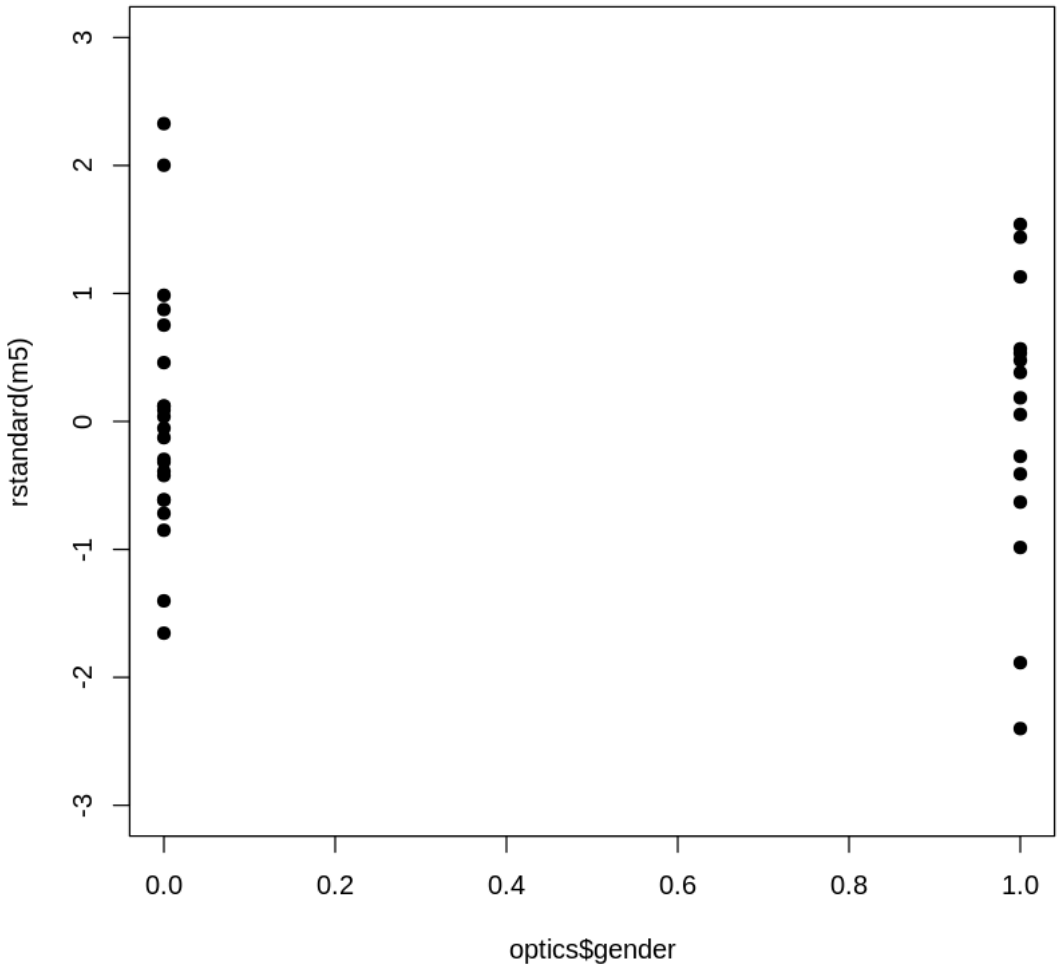
Further:

- Functional forms of model covariates: There is no evident linearity between the fitted values and residuals (plot 1 above). Thus, we may conclude that a linear model provides a decent fit to the data.
- Adequate fit of the covariates in the model: There are no noticable patterns in the plotted standardized residuals against covariates in the model (below), and the values seem to be decently clustered around the mean. Threfore, these covariates adequately fit the model. The additional covariate not included in the model also does not display a noticable pattern in the distribution of residuals, but does not seem to provide any relevant information that ought to be further explored. Still, a linear model seems to be the best fit for the data.

[15]:
```
#Standardized residuals vs covariates IN the model
plot(optics$OptPre, rstandard(m5), pch=19, ylim=c(-3, 3))
plot(optics$method_bin, rstandard(m5), pch=19, ylim=c(-3, 3))

#Standardized residuals vs covariates NOT IN the model
plot(optics$gender, rstandard(m5), pch=19, ylim=c(-3, 3))
```




3
2
1
0
-1
-2
-3
y
0
1
x



f) Predictive Accuracy: To assess the predictive accuracy of the chosen model (model 5), we may appeal to the standard error, which provides an estimate of how close model predictions are to observations (or how far observations fall from the fitted/regression line) in the units of the outcome variable (in this case OptPost score). In other words, it is the standard deviation of the error term in our model. Root Mean Squared Error (RMSE) and Residual Standard Error (RSE) are two measures of standard error, which only differ in that the latter is unbiased. That is, while RMSE uses a mean of squared residuals and hence, devides the sum by the sample size, RSE divides this sum by the degrees of freedom (the sample size minus the number of variables in the model), and is thus unbiased (despite being slightly larger). The formulas for RMSE and RSE are as follows:

$$RMSE = \sqrt{\frac{\Sigma_{i=1}^{N}(y_i - \hat{y_i})^2}{N}}$$

$$RSE = \sqrt{\frac{\Sigma_{i=1}^{N}(y_i - \hat{y}_i)^2}{df}}$$

The residual standard error (RSE) can be obtained from the summary table of the linear model (lm5) above, but is also calculated in two ways below. This value prodides us with an unbiased estimate of how "accurate" our model is. Thus, it provides an adequate measure of the predictive power of the model. Given the standard error values computed below, we may infer that this model predicts values that are on average 13-14 OptPost score units away from their true observations. That is, given one's OptPre score and the instructional method they recieved, the model will predict an OptPost score that is approximately 13-14 units away from the actual OptPost score. Moreover, from the summary tables above, we may notice that the RSE for saturated model is 13.6, which is very close to the RSE of the more parsimonious model of choice (13.91). Therefore, given our measure of predictive capability, our chosen model has almost the same level of predictive accuray as the saturated model, yet only contains 2 covariates and 3 terms, as opposed to 3 covariates and 8 terms like the saturated model. To test this predictive power, it may be wise to use additional test data. That is, we only know how well this model does at predicting OptPost scores on the training data provided. However, to see whether it has the level of predictive accuracy we expect, it serves us well to apply the it to new data.

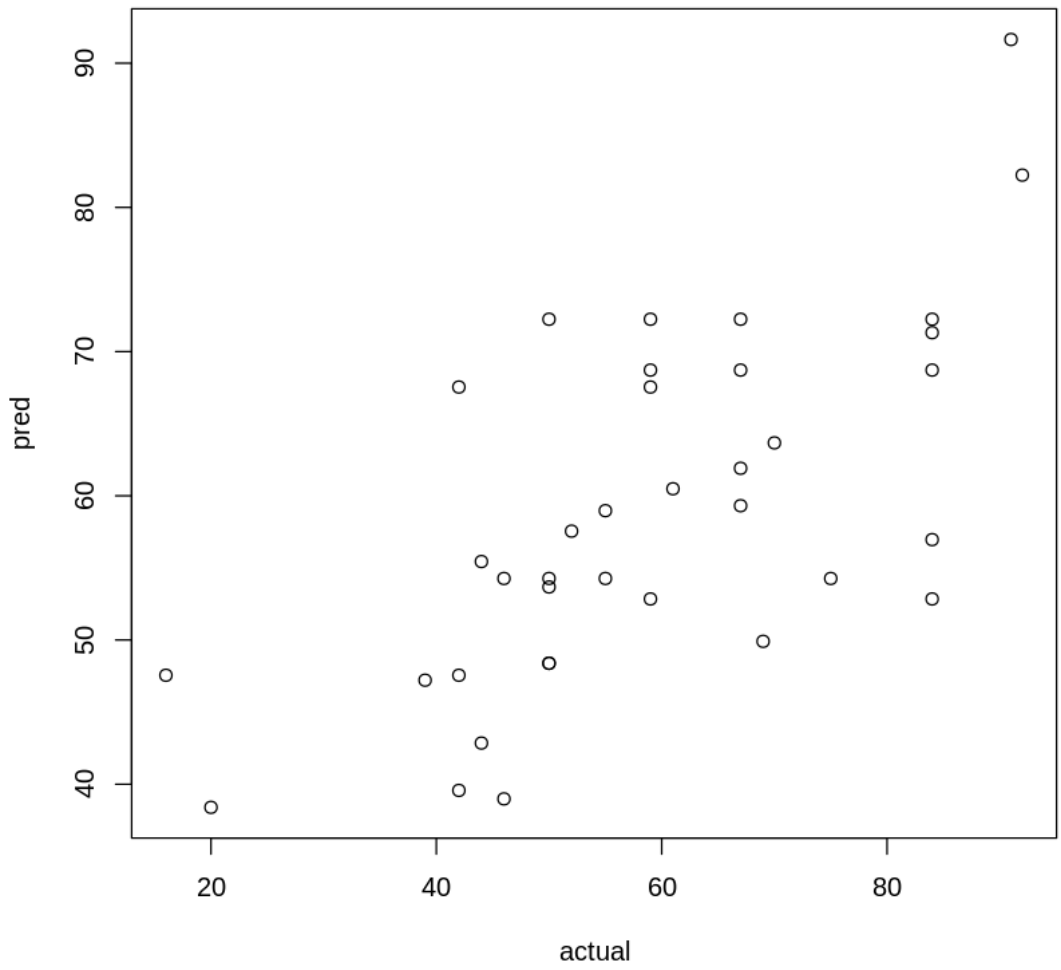
[16]:
```
#Predicted values vs. actual values
pva <- data.frame(actual = optics$OptPost, pred = predict(lm5))
plot(pva)

#predicted values, actual values, and residuals
pva_df <- pva %>% mutate(diff = actual - pred, sdiff = diff^2)
head(pva_df)

#largest and smallest residuals
max(abs(pva_df$diff))
min(abs(pva_df$diff))
```

A data.frame: 6 × 4

	actual <int>	pred <dbl>	diff <dbl>	sdiff <dbl>
1	50	72.24603	-22.2460253	494.8856436
2	67	72.24603	-5.2460253	27.5207819
3	61	60.48942	0.5105799	0.2606918
4	92	82.23914	9.7608602	95.2743922
5	59	67.54338	-8.5433833	72.9893974
6	16	47.55715	-31.5571544	995.8539927

31.5571543826153

0.510579876235262



[17]:
```
#RMSE is close to residual standard error (difference comes from dividing by␣
 ↪the df (sample size - #variables) instead of the sample size)
#mean divides by the sample size, so RMSE is more biased than residual standard␣
 ↪error (RSE)

#RMSE
sqrt(mean(pva_df$sdiff))

#RSE (this value is observed in the lm5 summary table above)
sqrt(sum(pva_df$sdiff)/(nrow(pva_df)-3))

#Different way of calculating RMSE and RSE
#RMSE
```

```r
mse <- mean(summary(lm5)$residuals^2)
rmse <- sqrt(mse)
rmse

#RSE
sqrt(deviance(lm5)/df.residual(lm5))
```

13.335948788449

13.9118639733703

13.335948788449

13.9118639733703

g) Interpretation of Regression Coefficients:

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 = 42.8545 + 0.5878 X_1 - 13.2761 X_2.$$

The chosen model (model 5) not only tells us that gender (X_3) and the relationship between OptPre score (X_1) and instruction method (X_2) are not significant in predicting OptPost scores (Y), but that, independently, OptPre score and instruction method have statistically significant predictive power over one's OptPost score. Particularly, it holds that for every 1-unit increase in OptPre score, we can expect an increase of about 0.59 in OptPost score, and moreover, that recieving a traditional instruction method results in an expected OptPost score that is approximately 13.28 units smaller than it would've been under the MBI instruction method (or is approximately 13.28 units smaller than the OptPost score of someone who had the same OptPre score, but recieved MBI instruction).

h) Results & Final Conclusion: Given the result of this analysis and the information obtained by the models, it is safe to assume that a linear relationship exists between OptPre scores, instruction method, and OptPost scores. Particularly, it is evident that the former two play a significant role in explaning some of the variation in the latter. Moreover, despite not knowing specifically the extent to which the instruction method influences OptPost scores, it is safe to conclude, based on the regression coefficients discussed previously, that the instruction method plays a much larger role in determining OptPost scores than does the OptPre score. Furthermore, the difference in predicted OptPre scores that results from the inclusion or exclusion of the MBI instruction method, allows us to infer that the newer instructuon method is more effective than the traditional one. Thus, we may deduce not only that there exists a strong association between instruction method and OptPost scores, but that the new instruction method (MBI) tends to produce overall higher OptPost scores than the traditional one, making it more preferable.

1.1.2 Question 2:

The dataset "Adelaide_grads.csv" presents the survival 50 years after graduation of men and women who graduated between 1938 and 1947 from various Faculties of the University of Adelaide (data compiled by J.A. Keats).

a) Perform a comprehensive EDA to inspect, understand and describe the information collected in this dataset. Use appropriate summary statistics and plots to present your results from the EDA. What do you observe?

Consider ONLY the information about the Faculties of Arts and Science (all years, both male and female) and answer the following questions:

b) Fit an appropriate model to answer the following questions:

- Are the proportions of graduates who survived for 50 years after graduation the same for all years of graduation?
- Are the proportions of graduates who survived for 50 years after graduation the same for males and females?
- Are the proportions of male graduates who survived for 50 years after graduation the same for Arts and Science?
- Are the proportions of female graduates who survived for 50 years after graduation the same for Arts and Science?
- Is the difference between males and females in the proportion of graduates who survived for 50 years after graduation the same for Arts and Science?

c) Choose a model selection procedure (backward, forward, or stepwise) to find the model that best fits your data. Check for all possible interactions terms among the covariates in the model. Which model best fits the data?

d) Assess the overall fit of the "best" model using regression diagnostics to identify problems indicating model inadequacy.

e) Use appropriate methods to assess the predictive accuracy of the "best" model.

f) Interpret all the regression coefficients of your "best" model.

Consider ONLY the information about male graduates (all years and Faculties) to answer the following questions:

g) You are interested in assessing the association between major and survivorship. Choose a model that best fits the data and answer this question using the respective regression coefficients.

h) Assess the overall fit of the model.

i) Compare the predictive accuracy of this model with that of the "best" resulting model in the previous part of Question 2.

[18]:
```
#DATA WRANGLING

#importing "Adelaide_grads" data
df2 <- read.csv("/home/jovyan/AGLM/HW2/Adelaide_grads.csv")
head(df2)
```

A data.frame: 6 × 5

	year <int>	survive <int>	total <int>	faculty <chr>	sex <chr>
1	1938	18	22	medicine	men
2	1939	16	23	medicine	men
3	1940	7	17	medicine	men
4	1941	12	25	medicine	men
5	1942	24	50	medicine	men
6	1943	16	21	medicine	men

[19]:
```
#WE HAVE MISSING VALUES (for year 1946)....OMITTING DATA FOR THAT YEAR
#(substituting with 0 will result in bias)
#removing year with missing values (1946)
df2 <- subset(df2, year != 1946)
rownames(df2) <- 1:nrow(df2)

head(df2)
nrow(df2)
unique(df2$faculty)
```

A data.frame: 6 × 5

	year <int>	survive <int>	total <int>	faculty <chr>	sex <chr>
1	1938	18	22	medicine	men
2	1939	16	23	medicine	men
3	1940	7	17	medicine	men
4	1941	12	25	medicine	men
5	1942	24	50	medicine	men
6	1943	16	21	medicine	men

54

1. 'medicine' 2. 'arts' 3. 'science' 4. 'engineering'

[20]:
```
#adding column for number of people who died for binary grouped data
agrads <- df2 %>% mutate(passed=total-survive)

#adding column for the proportion of those who survived (out of the total for␣
 ↪that given population) for EDA
agrads <- agrads %>% mutate(prop=survive/total)

#adding a column for categorical variable (year)
    #since it is ordinal, we can do this instead of creating 8 dummy variables
        #to create dummy variables we simply would convert year to factor:␣
 ↪data$col <- factor(data$col)
        #(this tells the model to treat it as categorical)

year <- c(1938, 1939, 1940, 1941, 1942, 1943, 1944, 1945, 1947)
yrs <- c(1, 2, 3, 4, 5, 6, 7, 8, 9)
yr_df <- data.frame(year, yrs)
agrads <- agrads %>% full_join(yr_df, by="year")
```

```
head(agrads)
```

A data.frame: 6 × 8

	year <dbl>	survive <int>	total <int>	faculty <chr>	sex <chr>	passed <int>	prop <dbl>	yrs <dbl>
1	1938	18	22	medicine	men	4	0.8181818	1
2	1939	16	23	medicine	men	7	0.6956522	2
3	1940	7	17	medicine	men	10	0.4117647	3
4	1941	12	25	medicine	men	13	0.4800000	4
5	1942	24	50	medicine	men	26	0.4800000	5
6	1943	16	21	medicine	men	5	0.7619048	6

a) Exploratory Data Analysis (EDA) The variables in this dataset are as follows: - Outcome (Y: graduates who survived after 50 years) - Covariate (X_1: year, X_2: faculty, X_3: sex)

The primary outcome of interest is a categorical binary random variable with nominal scale (survived or passed), while the predictor variables (covariates) are discrete/categorical with ordinal scale (X_1); categorical with nominal scale (X_2); and categorical/binary (X_3).

This EDA consists of: - Descriptive Statistics - Scatterplot - Boxplots - Bar Graphs

Descriptive Statistics:

- Summary of survival proportion (minimum value, 1st quartile, median, mean, 3rd quartile, maximum value) for the whole data and for each faculty group
- Summary of number of people who survived and number of people who passed (minimum value, 1st quartile, median, mean, 3rd quartile, maximum value) for the whole data and for each faculty group
- Standard deviation (SD) and variance of those who survived and those who passed
- Counting: Calculating the number of observations for each faculty group; the total number of men and women in the data and those who survived; as well as the total, the number survived, and the proportion of people alive in each faculty group

[21]:
```
#summary of survival proportion and survival proportion by faculty group and by␣
↪sex
summary(agrads$prop)
by(agrads$prop, agrads$faculty, summary, na.rm=TRUE)
by(agrads$prop, agrads$sex, summary, na.rm=TRUE)

#summary of survive and passed
summary(agrads$survive) #mean ~ 13
summary(agrads$passed) #mean ~ 5

#SD and variance of survive and passed
sd(agrads$survive)
var(agrads$survive)
sd(agrads$passed)
var(agrads$passed)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3333  0.6380  0.7584  0.7408  0.8534  1.0000
```

```
agrads$faculty: arts
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3333  0.5538  0.7009  0.6804  0.8182  1.0000
------------------------------------------------------------
agrads$faculty: engineering
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.5000  0.6364  0.7143  0.7073  0.7600  0.8889
------------------------------------------------------------
agrads$faculty: medicine
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.4118  0.4800  0.6957  0.6610  0.7619  0.8571
------------------------------------------------------------
agrads$faculty: science
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.6316  0.7714  0.8496  0.8579  1.0000  1.0000
```

```
agrads$sex: men
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3333  0.5614  0.7050  0.6669  0.7714  0.8889
------------------------------------------------------------
agrads$sex: women
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.6875  0.8182  0.8889  0.8887  1.0000  1.0000
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    8.00   12.00   12.81   16.00   32.00
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   2.000   5.000   5.259   7.000  26.000
```

6.96371749037927

48.4933612858141

4.69518093040644

22.0447239692523

[22]:
```
#COUNTING
#NOTE: There are 760 total men in the data set, while only 216 total women␣
 ↪(women make up about 22% of this data)

agrads %>% count(faculty) #there are only female populations for arts and␣
 ↪science
by(agrads$total, agrads$sex, sum)
by(agrads$survive, agrads$sex, sum) #510/760 ~ 67% men survived, 182/216 ~ 84%␣
 ↪women survived
```

```
#populations by faculty
agrads %>% group_by(faculty) %>%
  summarize(total = sum(total), survived = sum(survive), prop_survive =␣
 ↪survived/total)
```

A data.frame: 4 × 2

faculty	n
<chr>	<int>
arts	18
engineering	9
medicine	9
science	18

```
agrads$sex: men
[1] 760
------------------------------------------------------------
agrads$sex: women
[1] 216
```

```
agrads$sex: men
[1] 510
------------------------------------------------------------
agrads$sex: women
[1] 182
```

A tibble: 4 × 4

faculty	total	survived	prop_survive
<chr>	<int>	<int>	<dbl>
arts	324	217	0.6697531
engineering	142	103	0.7253521
medicine	241	155	0.6431535
science	269	217	0.8066914

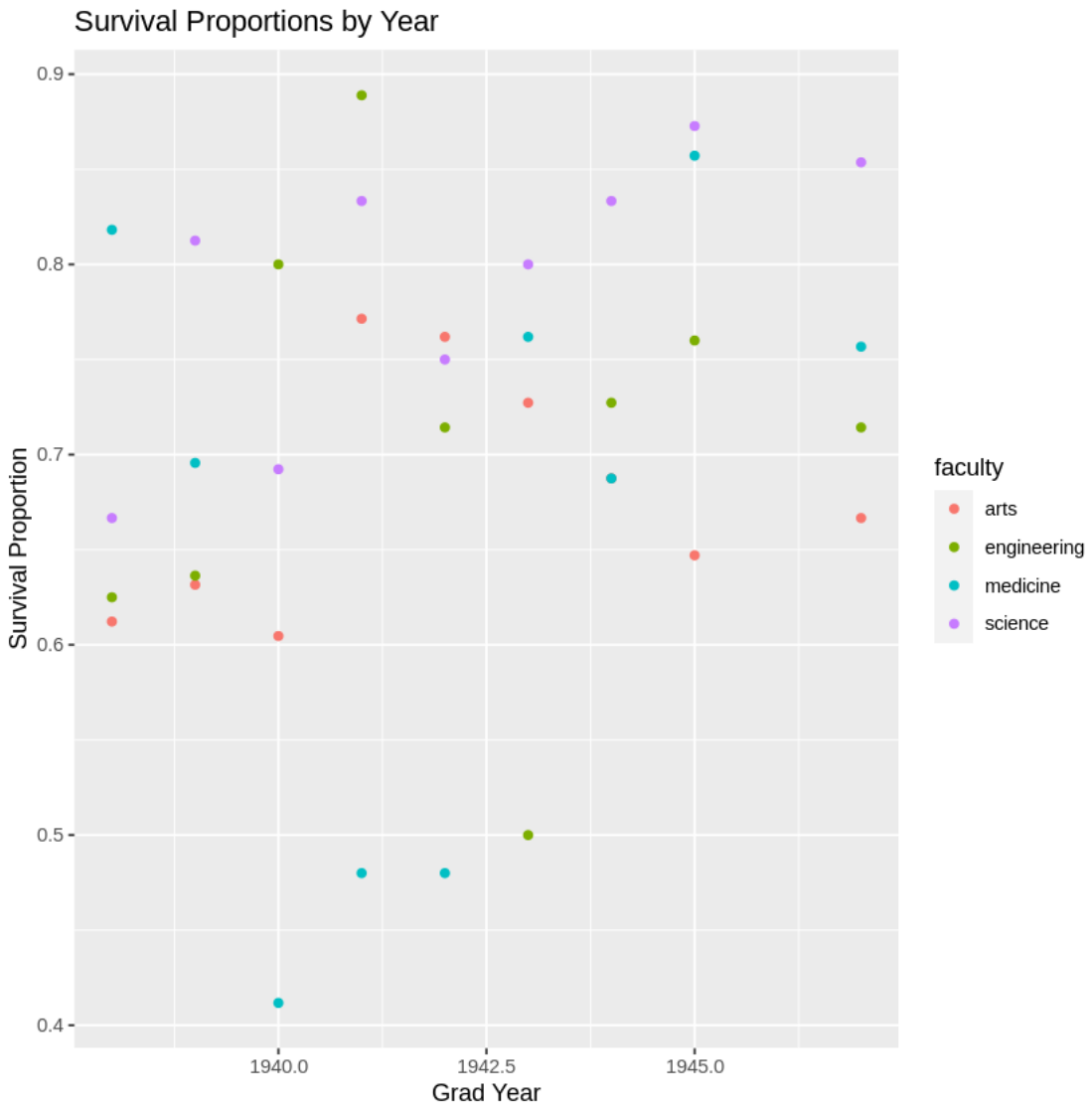
Graphs/Plots:

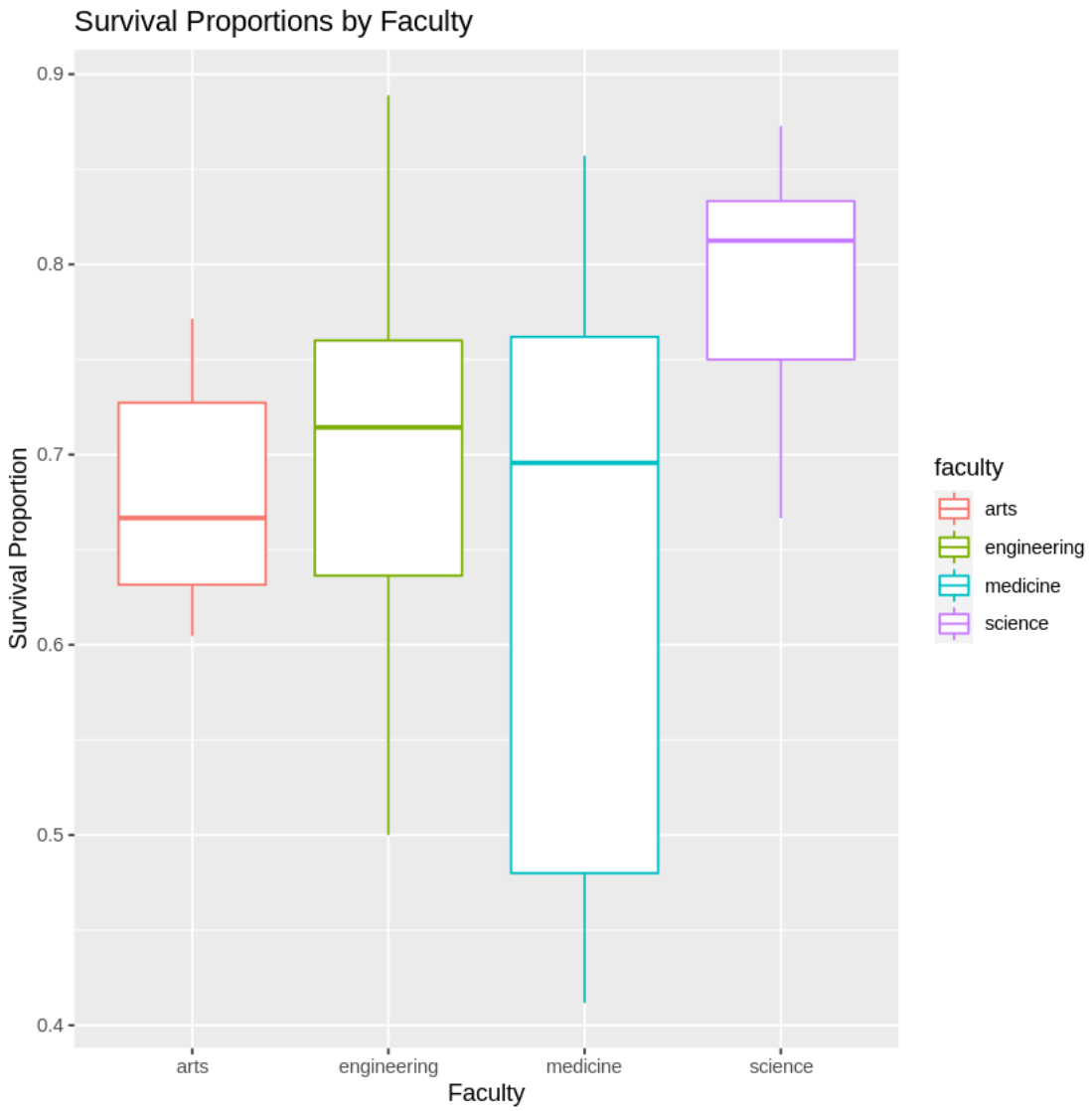
- Scatterplot (shows the relationship between graduation year and survival proportions for each faculty group)
- Boxplots (show a side by side comparison of the mean and spread of survival proportion for each faculty group)
- Bar graphs (show a side by side comparison of the change in survival proportions for men and women over time)

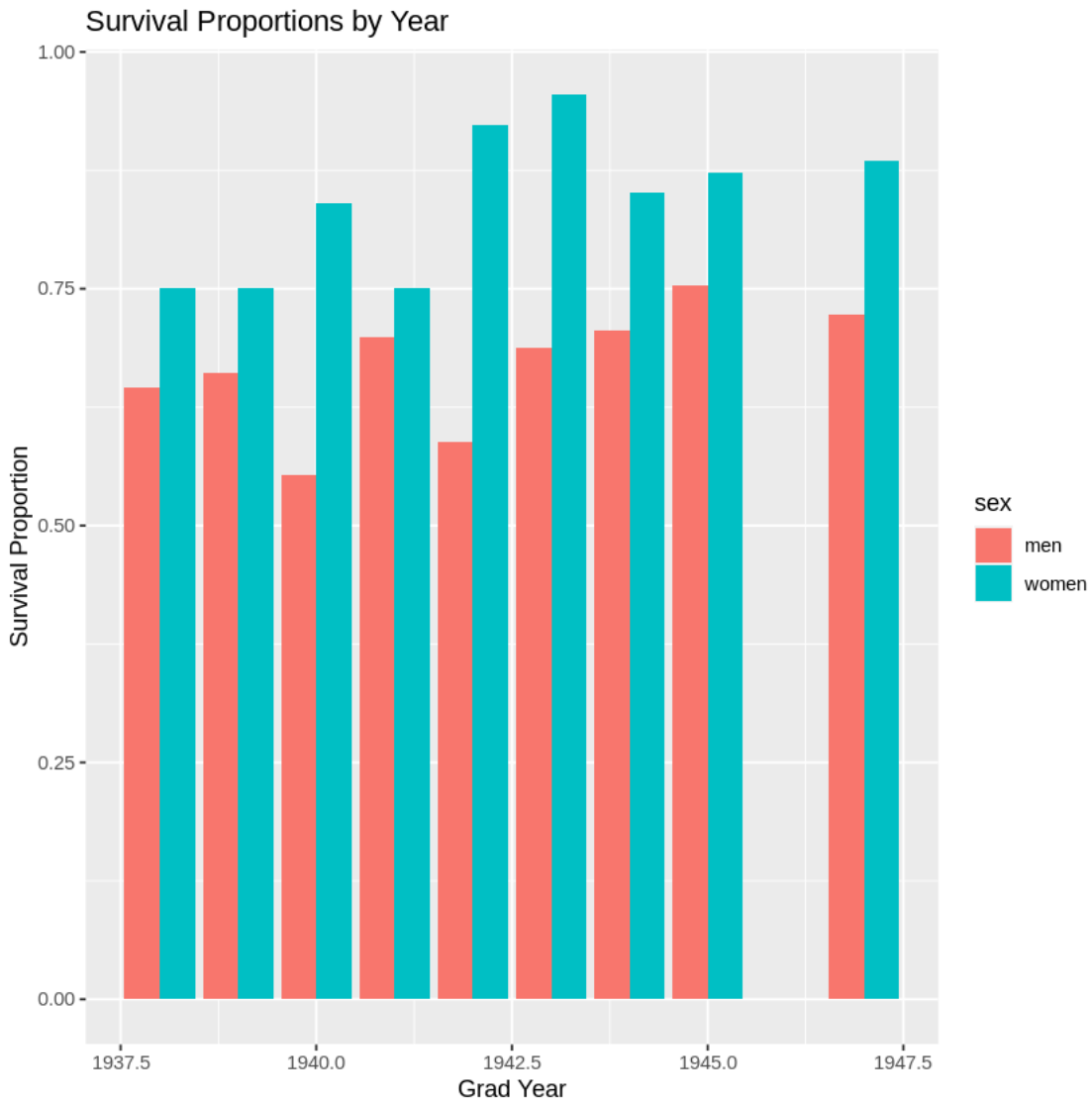
[23]:
```
#yearly survival proportions by faculty group
yearly_prop <- agrads %>%
    group_by(year, faculty) %>%
    summarize(total = sum(total), survived = sum(survive), prop_survive =␣
 ↪survived/total)

#yearly survival proportions by sex
yearly_prop1 <- agrads %>%
    group_by(year, sex) %>%
```

```
    summarize(total = sum(total), survived = sum(survive), prop_survive =␣
 ↪survived/total)

#Scatterplot
ggplot(yearly_prop) + geom_point(aes(x=year, y=prop_survive, color=faculty)) +
labs(x = "Grad Year", y = "Survival Proportion", title = "Survival Proportions␣
 ↪by Year")

#Boxplots
ggplot(yearly_prop, aes(group=faculty, x=faculty, y=prop_survive,␣
 ↪color=faculty)) + geom_boxplot() +
labs(x = "Faculty", y = "Survival Proportion", title = "Survival Proportions by␣
 ↪Faculty")

#Bar Graphs
ggplot(yearly_prop1, aes(x=year, y=prop_survive, fill=sex)) +␣
 ↪geom_bar(stat="identity", position=position_dodge()) +
labs(x = "Grad Year", y = "Survival Proportion", title = "Survival Proportions␣
 ↪by Year")
```

```
`summarise()` has grouped output by 'year'. You can override using the `.groups`
argument.

`summarise()` has grouped output by 'year'. You can override using the `.groups`
argument.
```


Survival Proportions by Year
Survival Proportion
0.9
0.8
0.7
0.6
0.5
0.4
1940.0
1942.5
1945.0
Grad Year
faculty
arts
engineering
medicine
science


Survival Proportions by Faculty
Survival Proportion
0.9
0.8
0.7
0.6
0.5
0.4
arts
engineering
medicine
science
Faculty
faculty
arts
engineering
medicine
science



b) Fitting a Model to Answer the Following: Considering ONLY information about the Faculties of Arts and Science (all years, both male and female). - Are the proportions of graduates who survived for 50 years after graduation the same for all years of graduation? - Given the model coefficients, there is no significant difference in proportions of graduates who survived for 50 years after graduation. The OR close to 1 tells us that proportions remained roughly the same for all years of graduation. - Are the proportions of graduates who survived for 50 years after graduation the same for males and females? - Given the model coefficients, the proportion of graduates who survived for 50 years after graduation is greater for females than for males. - Are the proportions of male graduates who survived for 50 years after graduation the same for Arts and Science? - Given the model coefficients, the proportion of male graduates who survived for 50 years after graduation is greater for Science than Arts. - Are the proportions of female graduates who survived for 50 years after graduation the same for Arts and Science? - Given the model coefficients, the proportion of female graduates who survived for 50 years after graduation is greater for Science than Arts. - Is the

difference between males and females in the proportion of graduates who survived for 50 years after graduation the same for Arts and Science? - Given the model coefficients, the difference between males and females in the proportion of graduates who survived for 50 years after graduation is slightly smaller for Science than Arts.

[24]:
```
#Data for arts and science faculty ONLY
as_grads = agrads %>% filter(faculty == "arts" | faculty == "science")
head(as_grads)
```

A data.frame: 6 × 8

	year <dbl>	survive <int>	total <int>	faculty <chr>	sex <chr>	passed <int>	prop <dbl>	yrs <dbl>
1	1938	16	30	arts	men	14	0.5333333	1
2	1939	13	22	arts	men	9	0.5909091	2
3	1940	11	25	arts	men	14	0.4400000	3
4	1941	12	14	arts	men	2	0.8571429	4
5	1942	8	12	arts	men	4	0.6666667	5
6	1943	11	20	arts	men	9	0.5500000	6

[25]:
```
#MODEL
#NOTE: When making a matrix for grouped binary data, the first of the two␣
 ↪columns should be the outcome of interest
#since we care about the proportion of grads who survived, this is the first␣
 ↪column of our matrix

asg_matrix <- as.matrix(as_grads[, c("survive", "passed")])
grads_glm1 <- glm(asg_matrix ~ as_grads$yrs + as_grads$faculty + as_grads$sex,␣
 ↪family=binomial(link="logit"))
summary(grads_glm1)
```

```
Call:
glm(formula = asg_matrix ~ as_grads$yrs + as_grads$faculty +
    as_grads$sex, family = binomial(link = "logit"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.73853  -0.44267   0.09422   0.73504   2.58393

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.07143    0.22285  -0.321    0.749
as_grads$yrs              0.05056    0.03693   1.369    0.171
as_grads$facultyscience   1.00314    0.21326   4.704 2.55e-06 ***
as_grads$sexwomen         1.27677    0.23053   5.538 3.05e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 80.903  on 35  degrees of freedom
Residual deviance: 28.684  on 32  degrees of freedom
AIC: 123.75

Number of Fisher Scoring iterations: 4
```

[26]:
```
#OR for covariates (exponentiating coefficients)
exp(coef(grads_glm1))
```

(Intercept) 0.93106350689074 **as_grads\$yrs** 1.05185648963447
as_grads\$facultyscience 2.72682955214678 **as_grads\$sexwomen** 3.58503190822848

Interpretation of Coefficients:

- Although year is not statistically significant, the model tells us that the odds of survival increase by a factor of $e^{0.05056} \approx 1.05$ for every unit increase in year.
- The odds of survival are $e^{1.00314} \approx 2.73$ greater for science faculty than for arts faculty.
- The odds of survival are $e^{1.27677} \approx 3.59$ greater for women than for men.

[27]:
```
#ORs and 95% CI
#Notice that the CI for years OR includes 1 (therefore it is not statistically␣
 ↪significant)
exp(cbind(OR = coef(grads_glm1), confint(grads_glm1)))
```

```
Waiting for profiling to be done…
```

A matrix: 4 × 3 of type dbl

	OR	2.5 %	97.5 %
(Intercept)	0.9310635	0.6012728	1.442597
as_grads$yrs	1.0518565	0.9784188	1.131051
as_grads$facultyscience	2.7268296	1.8040035	4.166391
as_grads$sexwomen	3.5850319	2.3048509	5.700151

[28]:
```
#Proportions of Survival
#using the inverse of logit (expit) to calculate survival proportions for each␣
 ↪scenario

exp(-0.07143+1.27677)/(1+exp(-0.07143+1.27677)) #females
exp(-0.07143)/(1+exp(-0.07143)) #males

exp(-0.07143+1.27677+1.00314)/(1+exp(-0.07143+1.27677+1.00314)) #females in␣
 ↪science
exp(-0.07143+1.27677)/(1+exp(-0.07143+1.27677)) #females in arts

exp(-0.07143+1.00314)/(1+exp(-0.07143+1.00314)) #males in science
exp(-0.07143)/(1+exp(-0.07143)) #males in arts
```

```
(exp(-0.07143+1.27677+1.00314)/(1+exp(-0.07143+1.27677+1.00314)))-(exp(-0.
↪07143+1.00314)/(1+exp(-0.07143+1.00314))) #females-males in science
(exp(-0.07143+1.27677)/(1+exp(-0.07143+1.27677)))-(exp(-0.07143)/(1+exp(-0.
↪07143))) #females-males in arts
```

0.769473377348014

0.48215008890617

0.901008437238705

0.769473377348014

0.717422078607233

0.48215008890617

0.183586358631471

0.287323288441844

c) Model Selection:

- Backward Selection

Based on the backward selection procedure, models 4 and 5 display the smallest AIC values of 208.56, and 210.01, respectively. Given that model 5 is the most parsimoneous of the two and the fact that the "year" variable is not statistically sigificant, it is safe to assume that model 5 fits and describes this data best:

$$g(E[Y]) = \beta_0 + \beta_{2e}X_{2e} + \beta_{2m}X_{2m} + \beta_{2s}X_{2s} + \beta_3 X_3$$

$$= 0.1560 + 0.8152X_{2e} + 0.4331X_{2m} + 1.0668X_{2s} + 1.2984X_3.$$

[29]:
```
#BWD selection

#matrix for grouped data
g_matrix <- as.matrix(agrads[, c("survive", "passed")])

#saturated model (all possible interaction terms)
g_glm1 <- glm(g_matrix ~ yrs*faculty*sex, family=binomial(link="logit"),␣
↪data=agrads)
summary(g_glm1)

#model 2 (only 2-way interaction terms included)
g_glm2 <- glm(g_matrix ~ yrs*faculty + yrs*sex + faculty*sex,␣
↪family=binomial(link="logit"), data=agrads)
summary(g_glm2)
```

```r
#model 3 (removing the least beneficial 2-way interaction term)
g_glm3 <- glm(g_matrix ~ yrs*faculty + yrs*sex, family=binomial(link="logit"),␣
 ↪data=agrads)
summary(g_glm3)

#model 4 (additive model with only the main effects)
g_glm4 <- glm(g_matrix ~ yrs + faculty + sex, family=binomial(link="logit"),␣
 ↪data=agrads)
summary(g_glm4)

#model 5 (removing the least beneficial covariate) <- BEST MODEL
g_glm5 <- glm(g_matrix ~ faculty + sex, family=binomial(link="logit"),␣
 ↪data=agrads)
summary(g_glm5)

#model 6 (null model)
g_glm6 <- glm(g_matrix ~ 1, family=binomial(link="logit"), data=agrads)
summary(g_glm6)
```

```
Call:
glm(formula = g_matrix ~ yrs * faculty * sex, family = binomial(link = "logit"),
    data = agrads)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3082  -0.4845   0.0521   0.6775   2.4523

Coefficients: (4 not defined because of singularities)
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                    0.31992    0.29740   1.076    0.282
yrs                           -0.02850    0.05600  -0.509    0.611
facultyengineering             0.49650    0.51292   0.968    0.333
facultymedicine               -0.09305    0.42836  -0.217    0.828
facultyscience                 0.17038    0.49446   0.345    0.730
sexwomen                       0.62198    0.50579   1.230    0.219
yrs:facultyengineering         0.05504    0.08545   0.644    0.520
yrs:facultymedicine            0.09838    0.07788   1.263    0.206
yrs:facultyscience             0.15119    0.08637   1.750    0.080 .
yrs:sexwomen                   0.12195    0.09491   1.285    0.199
facultyengineering:sexwomen         NA         NA      NA       NA
facultymedicine:sexwomen            NA         NA      NA       NA
facultyscience:sexwomen        1.15633    1.63211   0.708    0.479
yrs:facultyengineering:sexwomen     NA         NA      NA       NA
yrs:facultymedicine:sexwomen        NA         NA      NA       NA
yrs:facultyscience:sexwomen    -0.13195    0.26784  -0.493    0.622
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 113.89  on 53  degrees of freedom
Residual deviance:  48.88  on 42  degrees of freedom
AIC: 215.93

Number of Fisher Scoring iterations: 4



Call:
glm(formula = g_matrix ~ yrs * faculty + yrs * sex + faculty *
    sex, family = binomial(link = "logit"), data = agrads)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3082  -0.4853   0.0285   0.6851   2.4577

Coefficients: (2 not defined because of singularities)
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                 0.29359    0.29236   1.004   0.3153
yrs                        -0.02269    0.05474  -0.414   0.6786
facultyengineering          0.52283    0.51001   1.025   0.3053
facultymedicine            -0.06671    0.42488  -0.157   0.8752
facultyscience              0.24057    0.47448   0.507   0.6121
sexwomen                    0.69900    0.48278   1.448   0.1477
yrs:facultyengineering      0.04923    0.08463   0.582   0.5608
yrs:facultymedicine         0.09257    0.07698   1.203   0.2292
yrs:facultyscience          0.13736    0.08170   1.681   0.0927 .
yrs:sexwomen                0.10528    0.08856   1.189   0.2345
facultyengineering:sexwomen      NA         NA      NA       NA
facultymedicine:sexwomen         NA         NA      NA       NA
facultyscience:sexwomen     0.45446    0.67125   0.677   0.4984
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 113.891  on 53  degrees of freedom
Residual deviance:  49.125  on 43  degrees of freedom
AIC: 214.17

Number of Fisher Scoring iterations: 4



Call:
```

```
glm(formula = g_matrix ~ yrs * faculty + yrs * sex, family = binomial(link =␣
 ↪"logit"),
    data = agrads)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3082  -0.4929   0.0493   0.7432   2.5034

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             0.27557    0.28961   0.952   0.3413
yrs                    -0.02471    0.05428  -0.455   0.6489
facultyengineering      0.54086    0.50844   1.064   0.2874
facultymedicine        -0.04869    0.42299  -0.115   0.9084
facultyscience          0.29984    0.46275   0.648   0.5170
sexwomen                0.75047    0.47329   1.586   0.1128
yrs:facultyengineering  0.05125    0.08434   0.608   0.5434
yrs:facultymedicine     0.09460    0.07666   1.234   0.2172
yrs:facultyscience      0.13750    0.08075   1.703   0.0886 .
yrs:sexwomen            0.11011    0.08729   1.262   0.2071
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 113.89  on 53  degrees of freedom
Residual deviance:  49.62  on 44  degrees of freedom
AIC: 212.67

Number of Fisher Scoring iterations: 4



Call:
glm(formula = g_matrix ~ yrs + faculty + sex, family = binomial(link = "logit"),
    data = agrads)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.31717  -0.47299   0.09366   0.80456   2.58446

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.07426    0.19375  -0.383  0.70152
yrs                 0.05119    0.02754   1.859  0.06308 .
facultyengineering  0.74931    0.24285   3.086  0.00203 **
facultymedicine     0.39719    0.20210   1.965  0.04937 *
facultyscience      1.00238    0.21120   4.746 2.07e-06 ***
```

```
sexwomen            1.27655    0.23038   5.541 3.01e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 113.891  on 53  degrees of freedom
Residual deviance:  53.509  on 48  degrees of freedom
AIC: 208.56

Number of Fisher Scoring iterations: 4



Call:
glm(formula = g_matrix ~ faculty + sex, family = binomial(link = "logit"),
    data = agrads)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3526  -0.6661   0.1745   0.6781   2.5398

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.1560     0.1491   1.046 0.295705
facultyengineering   0.8152     0.2400   3.397 0.000681 ***
facultymedicine      0.4331     0.2008   2.157 0.031006 *
facultyscience       1.0668     0.2082   5.123 3.01e-07 ***
sexwomen             1.2984     0.2298   5.649 1.61e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 113.891  on 53  degrees of freedom
Residual deviance:  56.965  on 49  degrees of freedom
AIC: 210.01

Number of Fisher Scoring iterations: 4



Call:
glm(formula = g_matrix ~ 1, family = binomial(link = "logit"),
    data = agrads)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3860  -0.5305   0.5443   1.2431   2.9901
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.89061    0.07047   12.64   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 113.89  on 53  degrees of freedom
Residual deviance: 113.89  on 53  degrees of freedom
AIC: 258.94

Number of Fisher Scoring iterations: 4
```

d) Checking Model Fit: Assessing the overall fit of the model: - **Deviance**: LRT between chosen model and saturated model - When comparing then chosen model (5) with the saturated model, we obtain a p-value greater than 0.05. This indicates that we may not reject the null hypothesis, which states that both models fit the data equally well (notice they have very close deviances). Therefore, we may assume that our model is a good fit for the data, yet more parsimoneous than the saturated model. - Similarly, obtaining a significant p-value in the LRT between the chosen model (5) and the null model indicates that both do not fit the data equally well. As the null model is the worst fit for the data, we may conclude that conversely, the chosen model is a good fit.

Assessing the fit of particular observations (Regression Diagnostics): - **Outliers**: Standardized Residuals ~ N(0, 1) - In the plots below we observe that the standardized residuals are normally distributed (Q-Q plot), and that their absolute values do not exceed 3. Thus, we have no outliers in the data. - **Influential Points**: Cook's distance < 1, dfbetas $< 1 \sim$ N(0, 1) - In the plots below we observe that the Cook's distance of each value is less than 1 and that there are no dfbetas greater than 1. Thus, we have no influential points in the data.

[30]:
```r
#DEVIANCE

#LRT (same as Chisq)
#NULL: both fit the data equally well
#p>0.05 (when comparing model 5 with the saturated model)
#therefore, we do not reject the null and assume that our model is a good fit␣
 ↪for the data

#LRT for model 5 and saturated model
anova(g_glm5, g_glm1, test="LRT") #anova(g_glm4, g_glm1, test="Chisq")

#LRT for model 5 and null model (should have p<0.05)
anova(g_glm5, g_glm6, test="LRT")

#deviances are close
deviance(g_glm5)
```

```
deviance(g_glm1)
```

A anova: 2 × 5

	Resid. Df <dbl>	Resid. Dev <dbl>	Df <dbl>	Deviance <dbl>	Pr(>Chi) <dbl>
1	49	56.96460	NA	NA	NA
2	42	48.88015	7	8.08445	0.3252061

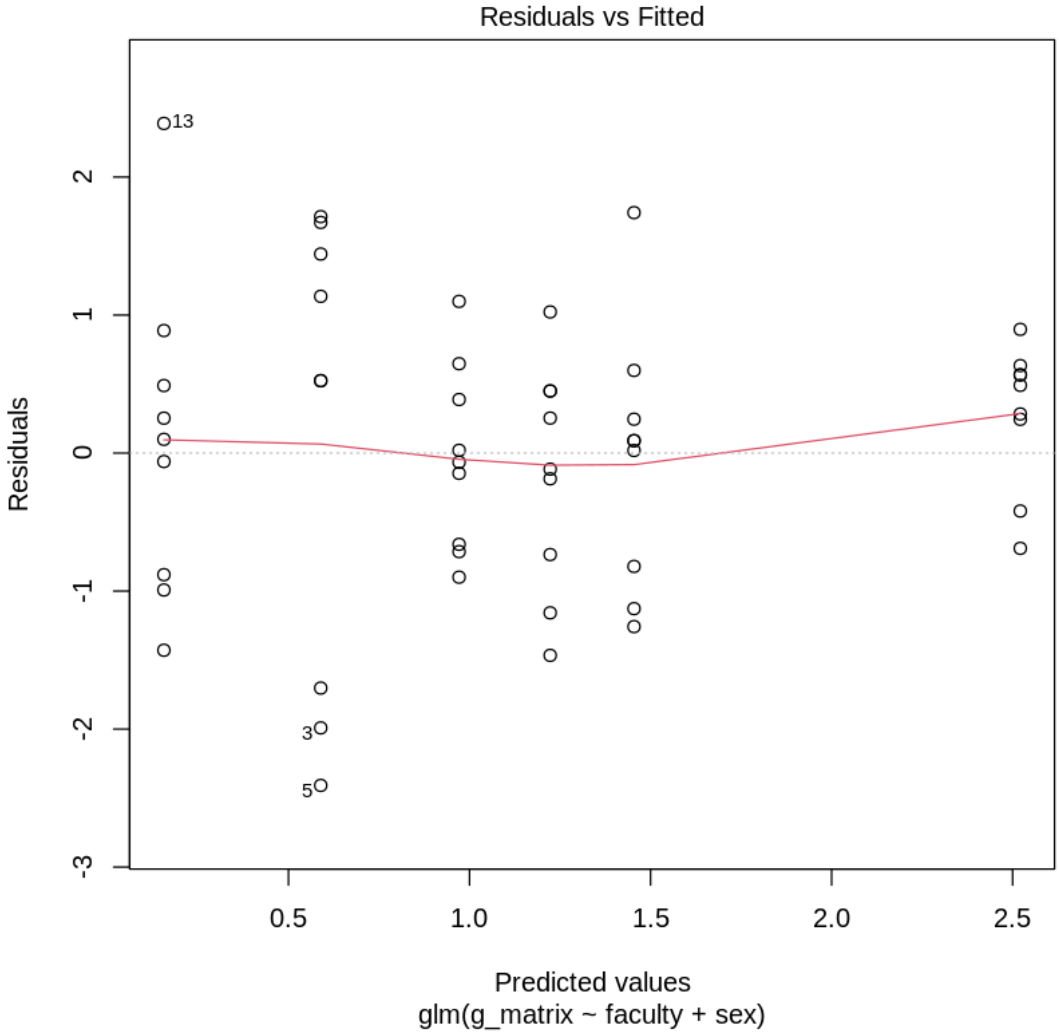
A anova: 2 × 5

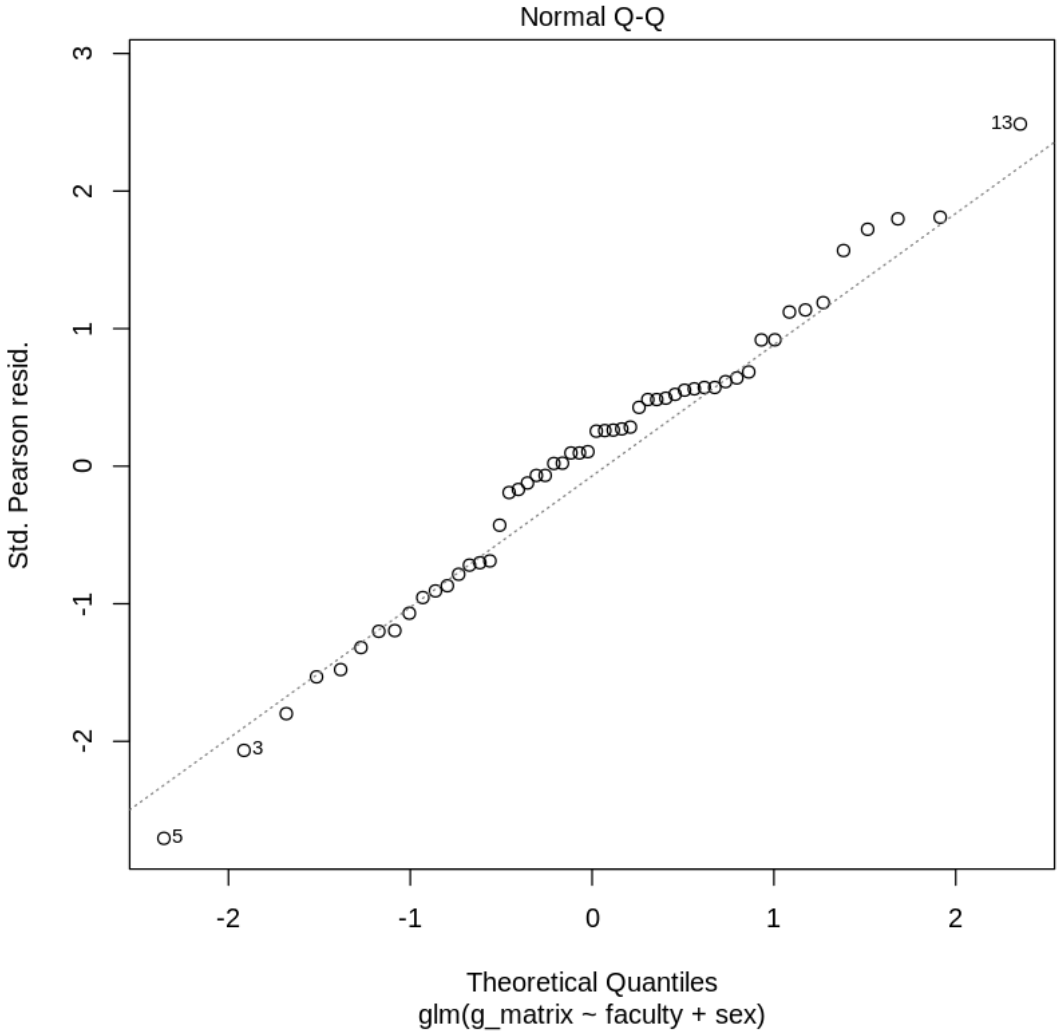
	Resid. Df <dbl>	Resid. Dev <dbl>	Df <dbl>	Deviance <dbl>	Pr(>Chi) <dbl>
1	49	56.9646	NA	NA	NA
2	53	113.8907	-4	-56.92608	1.282145e-11

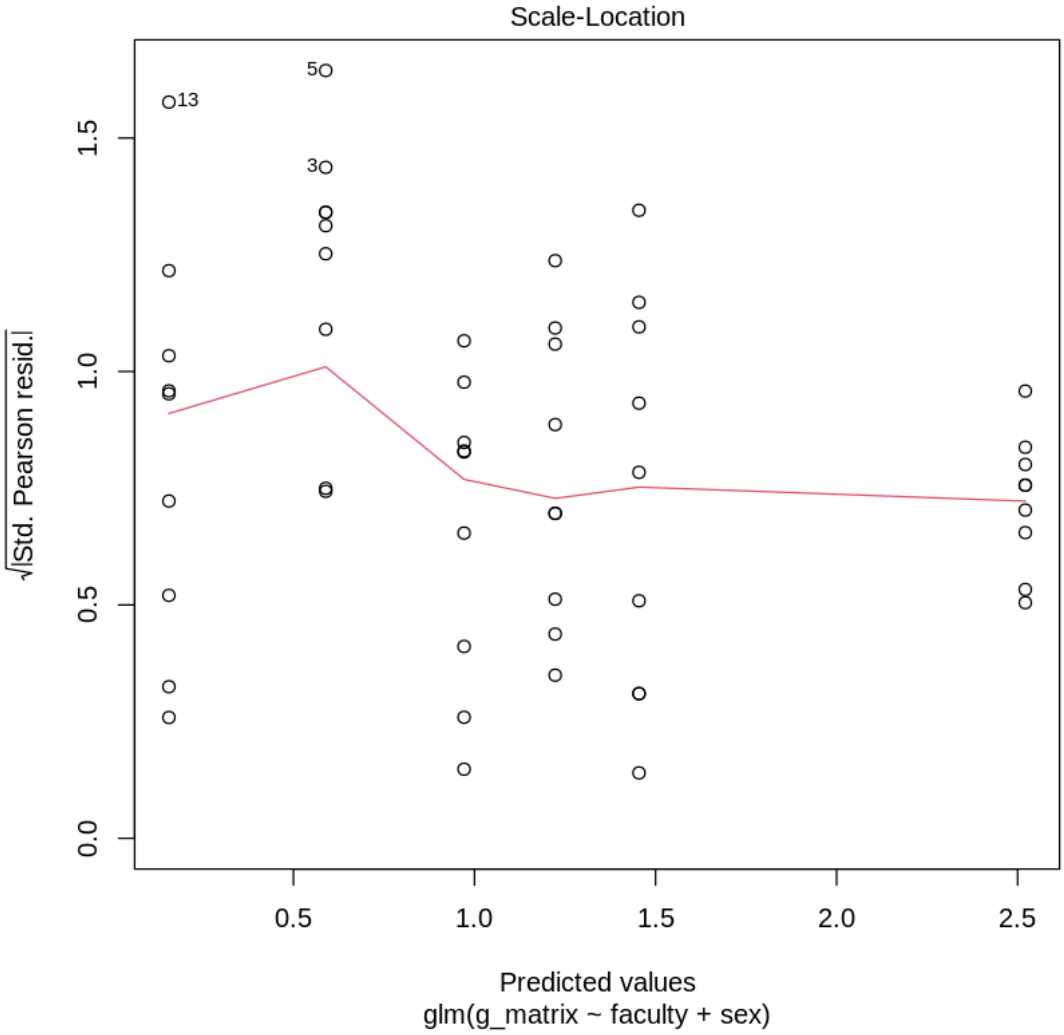
56.9645963125992

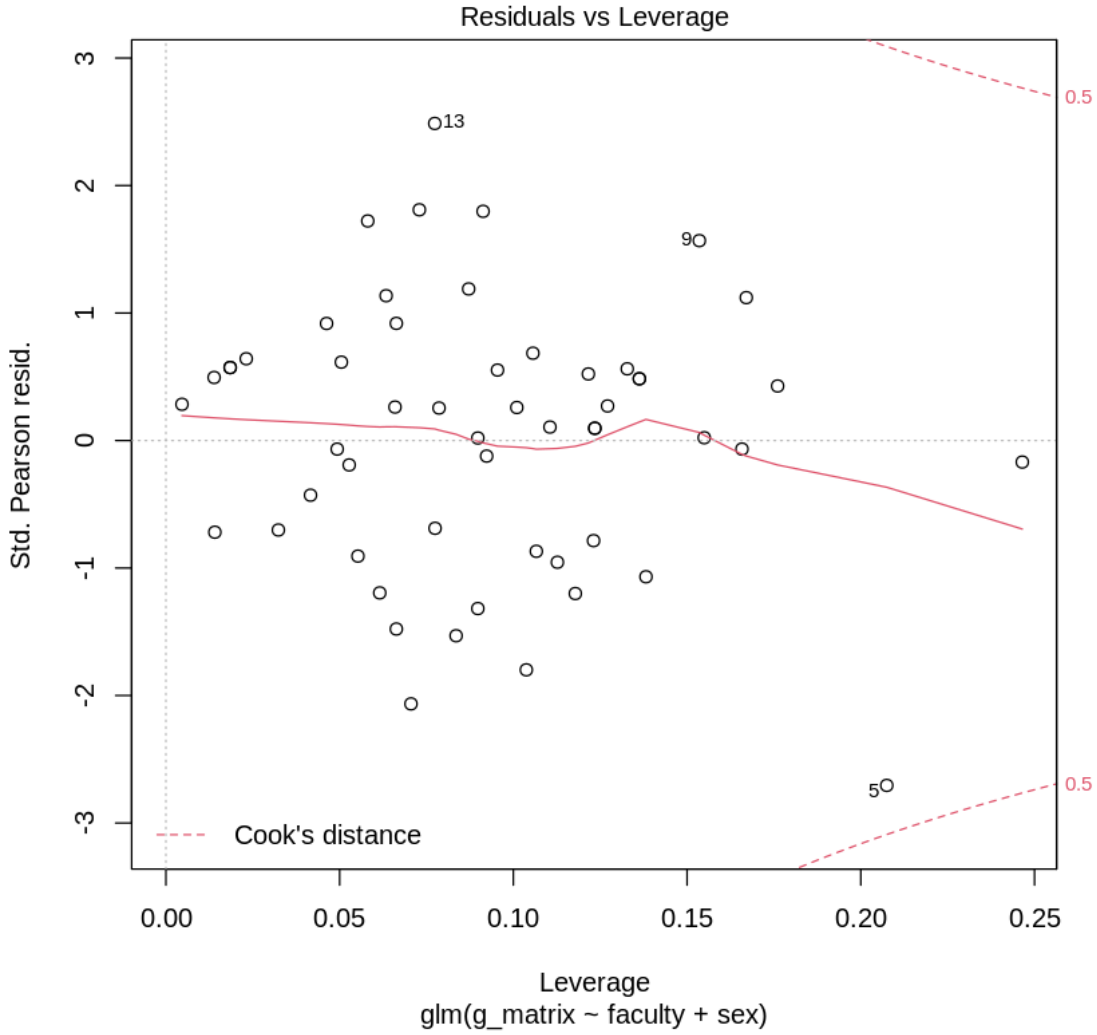
48.8801458788273

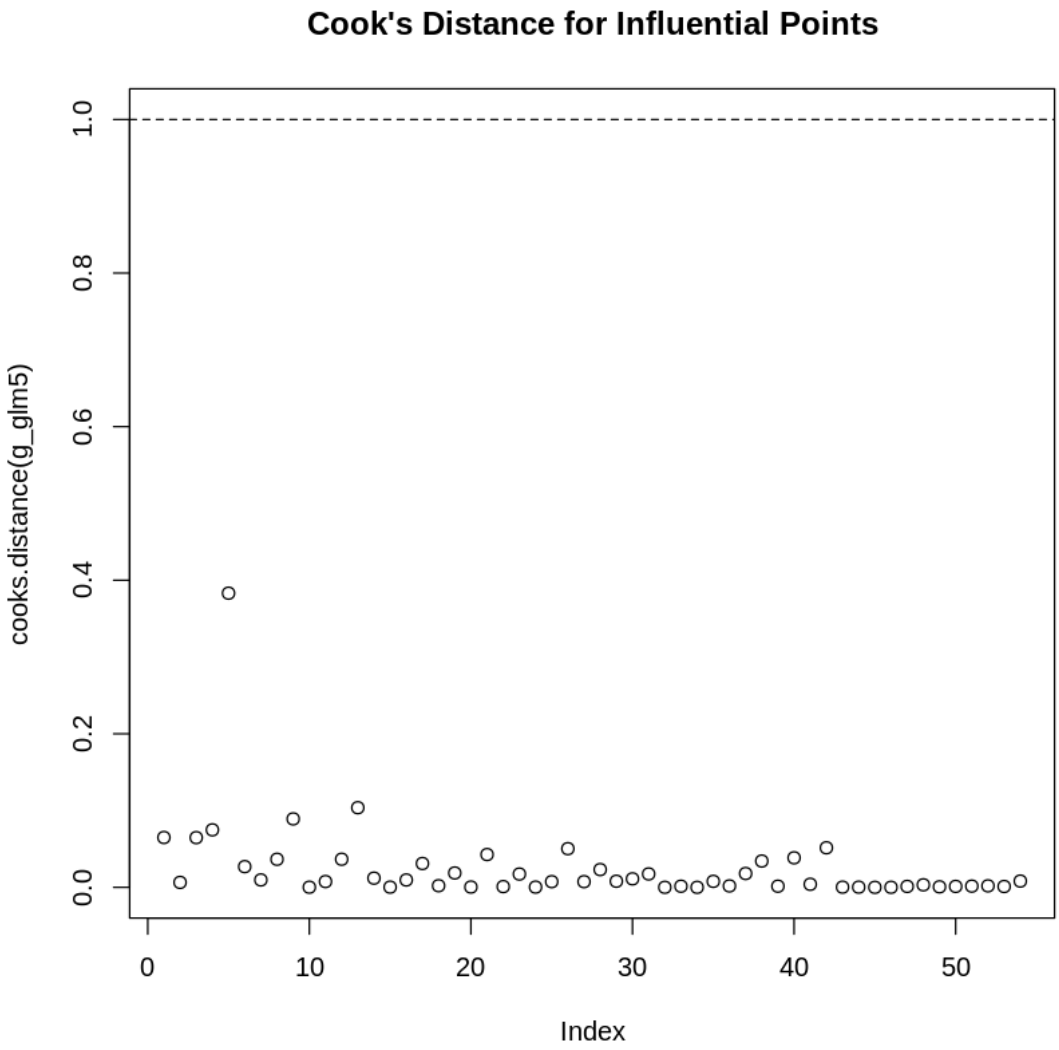
[31]:
```
#REGRESSION DIAGNOSTICS

#Outliers (plots 2 and 4)
#standardized residual are normally distributed (Q-Q plot)
#|standardized residuals| not greater than 3
#therefore, no outliers

#Influential Points (plot 5)
#Cook's distance is less than 1 (no values exceed the threshold)
#DFBETAS less than 1
#therefore, no influential points

plot(g_glm5)

plot(cooks.distance(g_glm5), ylim=c(0,1), main = "Cook's Distance for␣
 ↪Influential Points")
abline(h = 1, lty = 2) #cutoff line at 1 (degress of freedom/number of␣
 ↪observations is close to 1)

if (sum(abs(dfbetas(g_glm5))>1)) {print("dfbetas > 1")}
```


Residuals vs Fitted
Residuals
Predicted values
glm(g_matrix ~ faculty + sex)
13
3
5
2
1
0
-1
-2
-3
0.5
1.0
1.5
2.0
2.5


Normal Q-Q
Std. Pearson resid.
Theoretical Quantiles
glm(g_matrix ~ faculty + sex)
13
3
5
-2
-1
0
1
2
3


Scale-Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
13
5
3
0.5
1.0
1.5
2.0
2.5
Predicted values
glm(g_matrix ~ faculty + sex)


Residuals vs Leverage
Std. Pearson resid.
Leverage
glm(g_matrix ~ faculty + sex)
Cook's distance
0.5
0.5
13
9
5
3
2
1
0
-1
-2
-3
0.00
0.05
0.10
0.15
0.20
0.25



e) Predictive Accuracy/Power: Assessing the predictive accuracy of the chosen model: - Classification Table: - **ROC**: Graph of specificity against sensitivity ('fpr' against 'tpr') "illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied". "Classifiers that give curves closer to the top-left corner indicate a better performance. As a baseline, a random classifier is expected to give points lying along the diagonal (FPR = TPR). The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test". - The ROC below shows us that although our model's predictions are not random, they are not accurate enough to make the model reliable. - **AUC**: Area under the ROC "is equivalent to the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance". It tells us how much the model is capable of distinguishing between classes (equivalent to the concordance index or c-statistic). "The higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1". - In the plot below, we obtain an AUC of 0.6412, which means that there is a 64% chance that our model will be able to accurately predict a survival case versus a non-survival

case. This provides an overview of our model's predictive power. Given that an AUC of 0.5 would indicate that our model is 50% accurate and its predictions are purely random, it is safe to assume that it does not have very high predictive power. - **Correlation Measure**: $Corr(y, \hat{\mu})$ - Correlation coefficients of observed y's and predicted y's (from the chosen model), tell us the direction and strength of their linear relationship. A positive correlation close to 1 indicates a perfect linear relationship between observed and predicted values (that is, it idicates that the model is perfectly accurate). A correlation of 0 would indicate no linear relationship and suggest randomness in model predictions. - Having obtained a correlation coefficient of 0.2323 indicates that there is a weak positive linear relationship between observations in the data and our model's predicted values. This suggests that although a positive association does exists between observed and predicted values, it is not a very strong one. Thus, we can infer that our model isn't extremely successful at accurately predicting survival and non-survival cases, and hence, has minimal predictive power.

References: - https://en.wikipedia.org/wiki/Receiver_operating_characteristic - https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/ - https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

[32]:
```
#UNGROUPING DATA FOR PREDICTIVE ACCURACY

#ungrouped dataset
ug_agrads <- agrads %>%
   select(-c(passed, prop)) %>%
   uncount(total) %>%
   group_by(faculty, sex, year) %>%
   mutate(survive = as.integer(row_number() <= survive[1]))
head(ug_agrads)

#ungrouped glm for chosen model (5)
#coefficients should be the same as the grouped glm (below)
ug_glm <- glm(survive ~ faculty + sex, family=binomial(link="logit"),␣
 ↪data=ug_agrads)
summary(ug_glm)

#summary of grouped glm for chosen model (5):
#(Intercept)          0.1560     0.1491   1.046 0.295705
#facultyengineering   0.8152     0.2400   3.397 0.000681 ***
#facultymedicine      0.4331     0.2008   2.157 0.031006 *
#facultyscience       1.0668     0.2082   5.123 3.01e-07 ***
#sexwomen             1.2984     0.2298   5.649 1.61e-08 ***
```

976

```
Error in nrow(ug_agrads): object 'ug_agrads' not found
Traceback:

1. nrow(ug_agrads)
```

```
[ ]: #ROC and AUC
     #ROC: Specificity against Sensitivity ('fpr' against 'tpr')
     #AUC = Concordance Index C-statistic: Probability of accurate predictions
     library(pROC)

     multiclass.roc(ug_agrads$survive, ug_glm$fitted.values, plot=TRUE)
```

```
[ ]: #Correlation Measure
     #weak positive linear relationship
     cor(ug_agrads$survive, ug_glm$fitted.values)
```

f) Interpretation of Regression Coefficients:

$$g(E[Y]) = \beta_0 + \beta_{2e}X_{2e} + \beta_{2m}X_{2m} + \beta_{2s}X_{2s} + \beta_3 X_3$$

$$= 0.1560 + 0.8152X_{2e} + 0.4331X_{2m} + 1.0668X_{2s} + 1.2984X_3.$$

The chosen model (model 5) tells us that, independently, faculty and sex play a statistically significant role in explaining people's survival 50 years after graduating from University of Adelaide. Particularly, it holds that the odds of survival are $e^{1.2984} \approx 3.66$ greater for women than for men. Moreover, it suggests that, compared to the arts faculty (reference group), the odds of survival are $e^{0.8152} \approx 2.26$ greater for engineering faculty, $e^{0.4331} \approx 1.54$ greater for medicine faculty, and $e^{1.0668} \approx 2.91$ greater for science faculty.

```
[ ]: exp(1.2984) #female faculty
     exp(0.8152) #engineering faculty
     exp(0.4331) #medicine faculty
     exp(1.0668) #science faculty
```

Fitting a Model for the Following: Considering ONLY information about male graduates (all years and Faculties). - g) Assessing Variable Association: Major/Faculty and Survivorship - h) Checking Model Fit - i) Comparing Predictive Accuracy

```
[ ]: #Data for male faculty ONLY
     male_ug_agrads = ug_agrads %>% filter(sex == "men")

     #GLM for faculty and survivorship association
     male_ug_agrads$faculty <- factor(male_ug_agrads$faculty, levels=c("arts",␣
      ↪"engineering", "medicine", "science")) %>% #changing ordered factor to␣
      ↪ordinary factor
        relevel(male_ug_agrads$faculty, ref='arts') #specifying a reference group␣
      ↪for categorial variable 'faculty'
     male_glm <- glm(survive ~ faculty, family=binomial(link="logit"),␣
      ↪data=male_ug_agrads)
     summary(male_glm)
```

g) Assessing Variable Association: To assess the association between major/faculty and survivorship for male graduates, we fit the following glm:

$$g(E[Y]) = \beta_0 + \beta_{2e}X_{2e} + \beta_{2m}X_{2m} + \beta_{2s}X_{2s}$$

$$= 0.1911 + 0.7801X_{2e} + 0.3980X_{2m} + 0.9923X_{2s}.$$

The beta coefficients in this model tell us that, compared to the male arts faculty (reference group), the odds of survival are $e^{0.7801} \approx 2.18$ greater for male engineering faculty, $e^{0.3980} \approx 1.49$ greater for male medicine faculty, and $e^{0.9923} \approx 2.7$ greater for male science faculty. These odds ratios parallel the ones calculated for both male and female faculty. However, we can assume that values are greater when taking into account female faculty, as they have overall higher odds of suvival than males.

[ ]:
```
exp(0.7801) #engineering faculty
exp(0.3980) #medicine faculty
exp(0.9923) #science faculty
```

[ ]:
```
mf_or <- c(exp(0.8152), exp(0.4331), exp(1.0668))
m_or <- c(exp(0.7801), exp(0.3980), exp(0.9923))
ors <- data.frame(mf_or, m_or)
ors %>% mutate(diff = mf_or-m_or)
```

h) Checking Model Fit: Assessing the overall fit of the model: - **Deviance**: LRT between chosen model and null model - When comparing this model to the null model, we obtain a p-value less than 0.05. This indicates that both models do not fit the data equally well, and we may thus reject the null hypothesis. Given that the null model is the worst fit for the data, we may conclude that this model is a good fit.

Assessing the fit of particular observations (Regression Diagnostics): - **Outliers**: Standardized Residuals ~ N(0, 1) - In the first plot below we observe that the absolute values of standardized residuals do not exceed 3. Thus, there are no outliers in the data. - **Influential Points**: Cook's distance < 1, dfbetas < 1 ~ N(0, 1) - In the second plot below we observe that the Cook's distance of each value is less than 1 and that there are no dfbetas greater than 1. Thus, we have no influential points in the data.

[ ]:
```
#DEVIANCE

#LRT (same as Chisq)
#NULL: both fit the data equally well
#p<0.05
#therefore, we reject the null and assume that this model is a good fit for the␣
↪data

#null model
male_glm_null <- glm(survive ~ 1, family=binomial(link="logit"),␣
↪data=male_ug_agrads)
```

```
#LRT for model 5 and null model (should have p<0.05)
anova(male_glm, male_glm_null, test="LRT")
```

[ ]:
```
#REGRESSION DIAGNOSTICS

#Outliers (plot 1 - standardized residuals vs. fitted values)
#|standardized residuals| not greater than 3
#therefore, no outliers

plot(fitted(male_glm), rstandard(male_glm), pch=19, ylim=c(-3,3), main =␣
↪"Standardized Residuals vs. Fitted Values for Outliers")
abline(h = 0, lty = 2, lwd = 2)
abline(h = -3, lty = 3)
abline(h = 3, lty = 3)

#Influential Points (plot 2 - Cook's distance)
#Cook's distance is less than 1 (no values exceed the threshold)
#DFBETAS less than 1
#therefore, no influential points

plot(cooks.distance(male_glm), ylim=c(0,1), main = "Cook's Distance for␣
↪Influential Points")
abline(h = 1, lty = 2) #cutoff line at 1 (degress of freedom/number of␣
↪observations is close to 1)

if (sum(abs(dfbetas(male_glm))>1)) {print("dfbetas > 1")}
```

Model	AUC	Corr
Male & Female	0.6412	0.2323
Male	0.6017	0.1732

i) Comparing Predictive Accuracy: As expected, given the regression coefficients of this and the previous model, the predictive accuracy of the model including female faculty is slightly greater than that of this model, yet neither is very strong. While the AUC of the previous model tells us that there is a 64% chance it will accurately predict survival versus non-survival cases, the AUC of this model tells us that this probability decreases to 60% when considering only male faculty. Likewise, the correlation coefficient, which went from 0.2323 in the previous model to 0.1732 in this model, indicates that removing female faculty from the data results in an even weaker positive linear relationship between observations and predicted values. Although accounting for female faculty does increase the predictive accuracy of the model, it is difficult to determine whether sex truly plays a role in survivorship as we have inconsistent and limited data for females. That is, assessing the relationship between major/faculty and survivorship only in terms of male faculty removes a level of bias by providing more consistency within the data. Aside from this discrpency between the models however, both shed light on the extent to which major/faculty influences survivorship.

Given the limited predictive capabilities of the models, we may infer that although considering major/faculty will generally produce more accurate predictions than completely randomized ones, this predictor on its own is not sufficient to make our model truly reliable.

[ ]:
```
#PREDICTIVE ACCURACY
#ROC and AUC
#ROC: Specificity against Sensitivity ('fpr' against 'tpr')
#AUC = Concordance Index C-statistic: Probability of accurate predictions
multiclass.roc(male_ug_agrads$survive, male_glm$fitted.values, plot=TRUE)
```

[ ]:
```
#Correlation Measure
#weak positive linear relationship
cor(male_ug_agrads$survive, male_glm$fitted.values)
```

[ ]:
```
#COMPARISON
model <- c("Male & Female", "Male")
AUC <- c(0.6412,  0.6017)
Corr <- c(0.2323, 0.1732)
pa_comparison <- data.frame(model, AUC, Corr)
pa_comparison
```

1.1.3 Question 3:

The dataset "Beetle_mortality.csv" includes information from the Bliss (1935) study, about the numbers of beetles dead after five hours of exposure to gaseous carbon disulphide at various concentrations (example presented in class).

The primary objective is to evaluate the effect of dose on the beetles' mortality. For this purpose, use this dataset to fit GLMs for binary data. There is not restriction on the form (continuous, discrete, power transformation, number of groups, etc.) of the independent variable (dose) that you will include in the model. Just choose one form and work with it for answering the following questions:

a) Fit 3 different models using the following link functions:

- Logit
- Probit
- C-Log-Log

b) Interpret the regression coefficient estimate of the predictor in each of the 3 models.

[ ]:
```
#DATA WRANGLING

#importing "Beetle_mortality" data
beetles <- read.csv("/home/jovyan/AGLM/HW2/Beetle_mortality.csv")

#creating column for beetles alive
beetles <- beetles %>% mutate(alive = number-killed, prop_killed = killed/
 ↪number)
```

```r
head(beetles)
```

a) Model Fits for Different Link Functions:

- Logit
- Probit
- C-Log-Log

[ ]:
```r
#matrix for grouped data
b_matrix <- as.matrix(beetles[, c("killed", "alive")])

#LOGIT
b_logit <- glm(b_matrix ~ dose, family=binomial(link="logit"), data=beetles)
summary(b_logit)

#PROBIT
b_probit <- glm(b_matrix ~ dose, family=binomial(link="probit"), data=beetles)
summary(b_probit)

#C-LOG-LOG
b_cloglog <- glm(b_matrix ~ dose, family=binomial(link="cloglog"), data=beetles)
summary(b_cloglog)
```

[ ]:
```r
#MODEL COMPARISON

fits <- function(model){
    fit_p <- unique(fitted.values(model))
    fit_n <- as.vector(table(fitted.values(model)))
    fit_y <- fit_p*fit_n
    return(list("Fitted_Y"=fit_y, "Fitted_P"=fit_p))
}

#estimated number of beetles killed in each group
cbind(beetles$killed, fits(b_logit)[[1]], fits(b_probit)[[1]],␣
 ↪fits(b_cloglog)[[1]])

#estimated proportion of beetles killed in each group
cbind(beetles$killed, fits(b_logit)[[2]], fits(b_probit)[[2]],␣
 ↪fits(b_cloglog)[[2]])
```

[ ]:
```r
#Actual vs. Logit
plot(x=beetles$dose, y=beetles$prop_killed, frame=FALSE, type="b", pch=19,␣
 ↪col="black", xlab="Dose", ylab="Proportion Killed")
lines(x=beetles$dose, y=fits(b_logit)[[2]], type="b", pch=18, col="red", lty=2)
legend("topleft", legend=c("Actual", "Logit"), col=c("black", "red"), lty = 1:
 ↪2, cex=0.8)
```

```
[ ]: #Actual vs. Probit
plot(x=beetles$dose, y=beetles$prop_killed, frame=FALSE, type="b", pch=19,␣
 ↪col="black", xlab="Dose", ylab="Proportion Killed")
lines(x=beetles$dose, y=fits(b_probit)[[2]], type="b", pch=18, col="blue",␣
 ↪lty=2)
legend("topleft", legend=c("Actual", "Logit"), col=c("black", "blue"), lty = 1:
 ↪2, cex=0.8)
```

```
[ ]: #Actual vs. C-Log-Log
plot(x=beetles$dose, y=beetles$prop_killed, frame=FALSE, type="b", pch=19,␣
 ↪col="black", xlab="Dose", ylab="Proportion Killed")
lines(x=beetles$dose, y=fits(b_cloglog)[[2]], type="b", pch=18, col="green",␣
 ↪lty=2)
legend("topleft", legend=c("Actual", "Logit"), col=c("black", "green"), lty = 1:
 ↪2, cex=0.8)
```

b) Interpretation of Predictor Coefficients: Logit:

$$g(E[Y]) = \beta_0 + \beta_1 X_1 = -60.717 + 34.270 X_1$$

A beetle's odds of being killed increases by a factor of $e^{34.270}$ (`exp(34.270)`) for every one-unit increase in dose.

Probit:

$$g(E[Y]) = \beta_0 + \beta_1 X_1 = -34.935 + 19.728 X_1$$

A beetle's likelihood of being killed by a dose that is less than or equal to some dose, x, is given by `pnorm((19.728*x) - 34.935)`. Moreover, for a one-unit increase in dose, the z-score attributed to dose in the standard normal curve increases by 19.728.

C-Log-Log:

$$g(E[Y]) = \beta_0 + \beta_1 X_1 = -39.572 + 22.041 X_1$$

A beetle's probability of being killed by some dose, x, or $P(Y = 1|X = x)$, is given by $1 - e^{-e^{22.041x - 39.572}}$ (`1-(exp(-exp((22.041*x)-39.572)))`).

```
[ ]: min(beetles$dose)
max(beetles$dose)
```

```
[ ]: #INVERSE PROBIT
#(area under the standard normal curve to the left of x dose - cumulative␣
 ↪probabilities at different doses)

#Notice that we get almost 0 probability for the first 4 inputs because there␣
 ↪are not doses in the data
#(doses range from 1.6907 to 1.8839 - between the last two, that is why we go␣
 ↪from ~0 to ~1)
pnorm((19.728*0) - 34.935) #not a dose in the model
```

```
pnorm((19.728*0.5) - 34.935) #not a dose in the model
pnorm((19.728*1) - 34.935) #not a dose in the model
pnorm((19.728*1.5) - 34.935) #not a dose in the model
pnorm((19.728*2) - 34.935) #not a dose in the model
```

[ ]:
```
#Cumulative probabilities for every 0.01 increase in dose
#Notice that probabilities range from ~0 to ~1
pnorm((19.728*1.68) - 34.935)
pnorm((19.728*1.69) - 34.935)
pnorm((19.728*1.7) - 34.935)
pnorm((19.728*1.71) - 34.935)
pnorm((19.728*1.72) - 34.935)
pnorm((19.728*1.73) - 34.935)
pnorm((19.728*1.74) - 34.935)
pnorm((19.728*1.75) - 34.935)
pnorm((19.728*1.76) - 34.935)
pnorm((19.728*1.77) - 34.935)
pnorm((19.728*1.78) - 34.935)
pnorm((19.728*1.79) - 34.935)
pnorm((19.728*1.8) - 34.935)
pnorm((19.728*1.81) - 34.935)
pnorm((19.728*1.82) - 34.935)
pnorm((19.728*1.83) - 34.935)
pnorm((19.728*1.84) - 34.935)
pnorm((19.728*1.85) - 34.935)
pnorm((19.728*1.86) - 34.935)
pnorm((19.728*1.87) - 34.935)
pnorm((19.728*1.88) - 34.935)
pnorm((19.728*1.89) - 34.935)
```

[ ]:
```
#Incremental differences in cumulative probabilities for every 0.01 increase in␣
↪dose
#Notice that these values resemble a normal curve
#Imagining these differences as small slices along the curve with equal width␣
↪(x)
#We can see why the their areas would be smaller at the tails and largest in␣
↪the middle
pnorm((19.728*1.69) - 34.935) - pnorm((19.728*1.68) - 34.935)
pnorm((19.728*1.7) - 34.935) - pnorm((19.728*1.69) - 34.935)
pnorm((19.728*1.71) - 34.935) - pnorm((19.728*1.7) - 34.935)
pnorm((19.728*1.72) - 34.935) - pnorm((19.728*1.71) - 34.935)
pnorm((19.728*1.73) - 34.935) - pnorm((19.728*1.72) - 34.935)
pnorm((19.728*1.74) - 34.935) - pnorm((19.728*1.73) - 34.935)
pnorm((19.728*1.75) - 34.935) - pnorm((19.728*1.74) - 34.935)
pnorm((19.728*1.76) - 34.935) - pnorm((19.728*1.75) - 34.935)
pnorm((19.728*1.77) - 34.935) - pnorm((19.728*1.76) - 34.935)
pnorm((19.728*1.78) - 34.935) - pnorm((19.728*1.77) - 34.935)
```

```
pnorm((19.728*1.79) - 34.935) - pnorm((19.728*1.78) - 34.935)
pnorm((19.728*1.8) - 34.935) - pnorm((19.728*1.79) - 34.935)
pnorm((19.728*1.81) - 34.935) - pnorm((19.728*1.8) - 34.935)
pnorm((19.728*1.82) - 34.935) - pnorm((19.728*1.81) - 34.935)
pnorm((19.728*1.83) - 34.935) - pnorm((19.728*1.82) - 34.935)
pnorm((19.728*1.84) - 34.935) - pnorm((19.728*1.83) - 34.935)
pnorm((19.728*1.85) - 34.935) - pnorm((19.728*1.84) - 34.935)
pnorm((19.728*1.86) - 34.935) - pnorm((19.728*1.85) - 34.935)
pnorm((19.728*1.87) - 34.935) - pnorm((19.728*1.86) - 34.935)
pnorm((19.728*1.88) - 34.935) - pnorm((19.728*1.87) - 34.935)
pnorm((19.728*1.89) - 34.935) - pnorm((19.728*1.88) - 34.935)
```

[ ]:
```
#PROBIT
#(z-scores on the standard normal curve attributed to x dose - how many␣
 ↪standard deviations x is from the mean)
((19.728*1.68) - 34.935)
((19.728*1.69) - 34.935)
((19.728*1.7) - 34.935)
((19.728*1.71) - 34.935)
((19.728*1.72) - 34.935)
((19.728*1.73) - 34.935)
((19.728*1.74) - 34.935)
((19.728*1.75) - 34.935)
((19.728*1.76) - 34.935)
((19.728*1.77) - 34.935)
((19.728*1.78) - 34.935)
((19.728*1.79) - 34.935)
((19.728*1.8) - 34.935)
((19.728*1.81) - 34.935)
((19.728*1.82) - 34.935)
((19.728*1.83) - 34.935)
((19.728*1.84) - 34.935)
((19.728*1.85) - 34.935)
((19.728*1.86) - 34.935)
((19.728*1.87) - 34.935)
((19.728*1.88) - 34.935)
((19.728*1.89) - 34.935)
```

[ ]:
```
#INVERSE C-LOG-LOG
#Gives P(Y=1|X=x)
1-(exp(-exp((22.041*1.69)-39.572)))
1-(exp(-exp((22.041*1.7)-39.572)))
1-(exp(-exp((22.041*1.71)-39.572)))
1-(exp(-exp((22.041*1.72)-39.572)))
1-(exp(-exp((22.041*1.73)-39.572)))
1-(exp(-exp((22.041*1.74)-39.572)))
1-(exp(-exp((22.041*1.75)-39.572)))
```

```
1-(exp(-exp((22.041*1.76)-39.572)))
1-(exp(-exp((22.041*1.77)-39.572)))
1-(exp(-exp((22.041*1.78)-39.572)))
1-(exp(-exp((22.041*1.79)-39.572)))
1-(exp(-exp((22.041*1.8)-39.572)))
1-(exp(-exp((22.041*1.81)-39.572)))
1-(exp(-exp((22.041*1.82)-39.572)))
1-(exp(-exp((22.041*1.83)-39.572)))
1-(exp(-exp((22.041*1.84)-39.572)))
1-(exp(-exp((22.041*1.85)-39.572)))
1-(exp(-exp((22.041*1.86)-39.572)))
1-(exp(-exp((22.041*1.87)-39.572)))
1-(exp(-exp((22.041*1.88)-39.572)))
1-(exp(-exp((22.041*1.89)-39.572)))
```