HW3_6120

Jacqui Unciano

2023-07-05

```
library(leaps)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
data = read.table("C:\\Users\\Jacqueline\\OneDrive\\Documents\\MSDS\\datasets\\nfl.txt", header = T)
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(faraway)
```

```
##
## Attaching package: 'faraway'

## The following objects are masked from 'package:car':
##
##     logit, vif
```

Question 1

(a) Use the regsubsets() function from the leaps package to run all possible regressions. Set nbest=1. Identify the model (the predictors and the corresponding estimated coefficients) that is best in terms of

i. Adjusted R2
ii. Mallow's Cp
iii. BIC

```
allreg = regsubsets(y~., data=data, nbest=1)
summary(allreg)
```

```
## Subset selection object
## Call: regsubsets.formula(y ~ ., data = data, nbest = 1)
## 9 Variables  (and intercept)
##    Forced in Forced out
## x1     FALSE      FALSE
## x2     FALSE      FALSE
## x3     FALSE      FALSE
## x4     FALSE      FALSE
## x5     FALSE      FALSE
## x6     FALSE      FALSE
## x7     FALSE      FALSE
## x8     FALSE      FALSE
## x9     FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          x1  x2  x3  x4  x5  x6  x7  x8  x9 
## 1  ( 1 ) " " " " " " " " " " " " " " "*" " "
## 2  ( 1 ) " " "*" " " " " " " " " " " "*" " "
## 3  ( 1 ) " " "*" " " " " " " " " "*" "*" " "
## 4  ( 1 ) " " "*" " " " " " " " " "*" "*" "*"
## 5  ( 1 ) "*" "*" " " " " " " " " "*" "*" "*"
## 6  ( 1 ) " " "*" "*" "*" " " " " "*" "*" "*"
## 7  ( 1 ) " " "*" "*" "*" " " "*" "*" "*" "*"
## 8  ( 1 ) "*" "*" "*" "*" " " "*" "*" "*" "*"
```

```
which.max(summary(allreg)$adjr2)
```

```
## [1] 4
```

```
which.min(summary(allreg)$cp)
```

```
## [1] 3
```

```
which.min(summary(allreg)$bic)
```

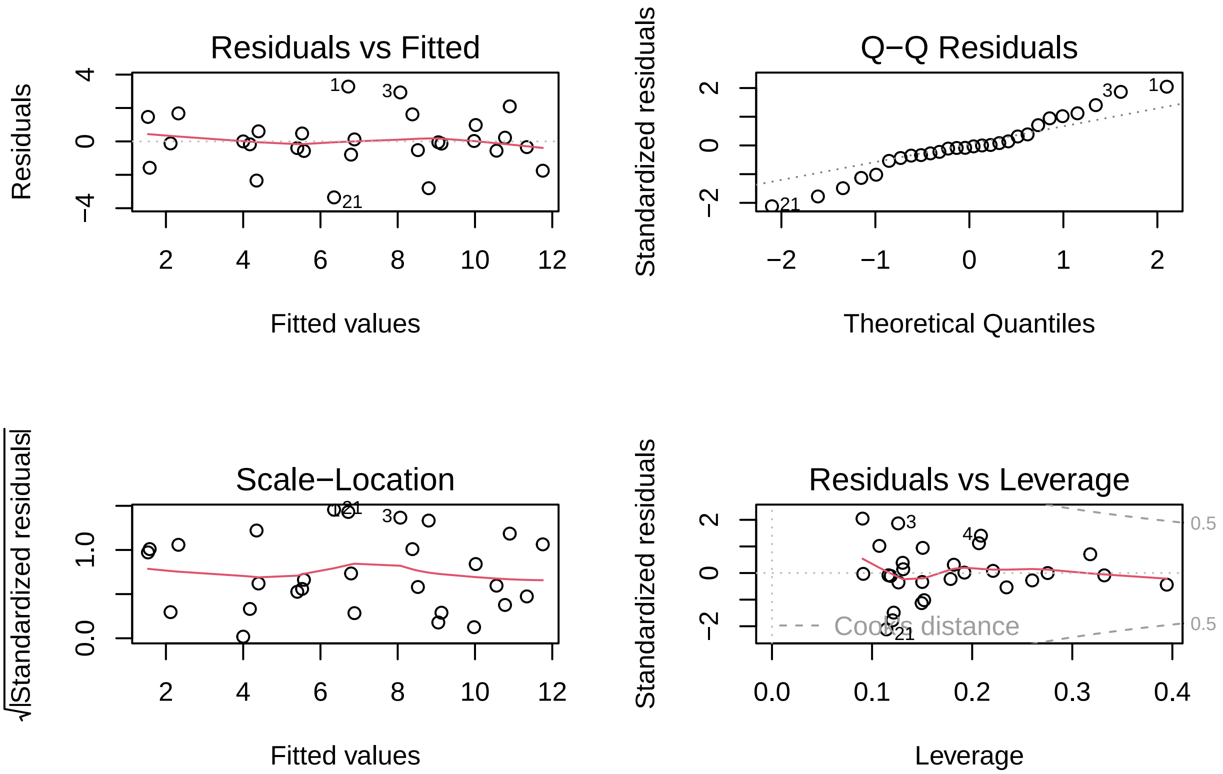
```
## [1] 3
```

The model chosen by all possible regression that is best in terms of ajdusted R2 is the model with x2, x7, x8, and x9.

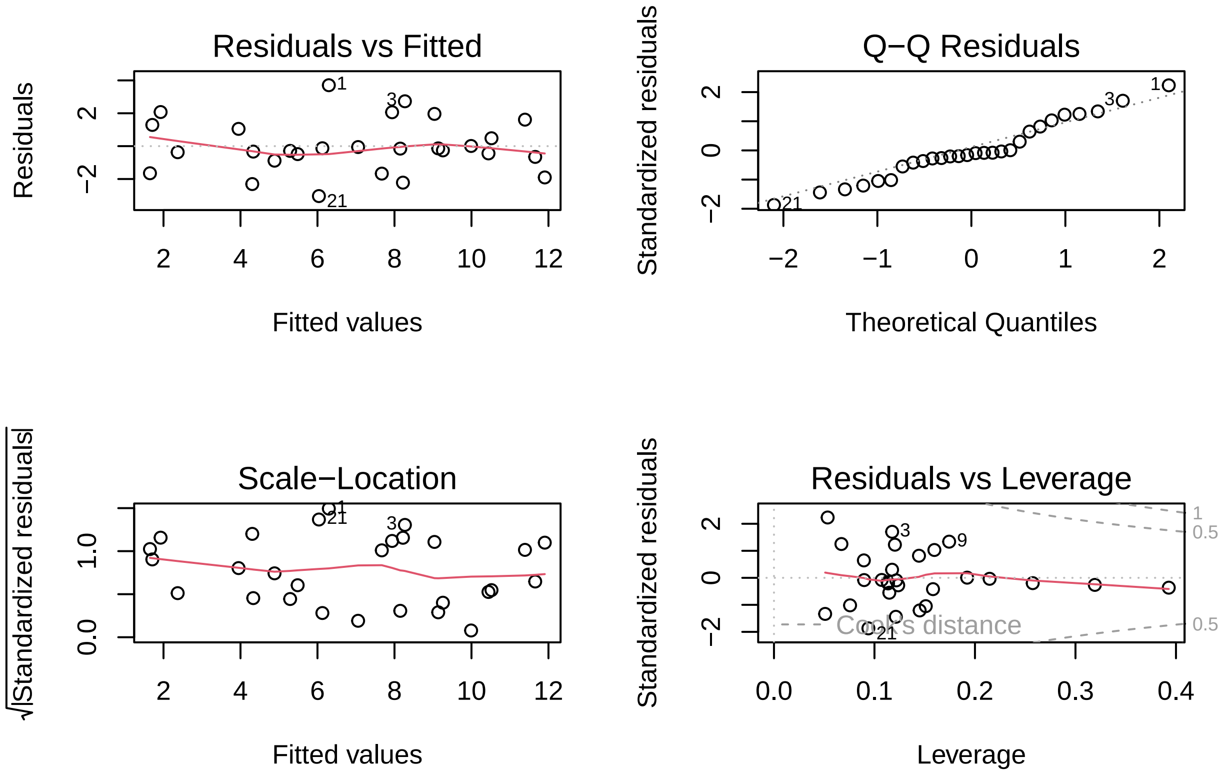
The model chosen by all possible regression that is best in terms of Mallow's CP and BIC is the model with x2, x7, and x8.

(b) For the models found in part 1a, use residual plots to assess if the regression assumptions are met, and address if any variables need to be transformed. If needed, transform the appropriate variable, and re-do part 1a using the transformed variables.

```
mod1 = lm(y~x2+x7+x8+x9, data=data)
mod2 = lm(y~x2+x7+x8, data=data)
par(mfrow=c(2,2))
plot(mod1)
```



```
plot(mod2)
```



Yes, the regression assumptions are met. The data points are distributed evenly over both the horizontal and vertical band. Meaning the mean error of residuals for each fitted value is about 0 and the variance is constant as the fitted values increase.

We could argue that the normality assumption is violated because of the QQ plot in model 2, however, I would say the data is fairly robust and allows for it, especially since the data isn't going to be perfect anyways.

(c) Run forward selection, starting with an intercept-only model. Report the predictors and the estimated coefficients of the model selected.

```
reginter = lm(y~1, data=data)
regfull = lm(y~., data=data)
step(reginter, scope=list(lower=reginter, upper=regfull), direction="forward")
```

```
## Start:  AIC=70.81
## y ~ 1
##
##          Df Sum of Sq    RSS    AIC
## + x8      1   178.092 148.87 50.785
## + x1      1   115.068 211.90 60.669
## + x7      1    97.238 229.73 62.931
## + x5      1    86.116 240.85 64.255
## + x2      1    76.193 250.77 65.385
## + x9      1    30.167 296.80 70.104
## <none>                326.96 70.814
## + x4    1    21.844 305.12 70.878
## + x6    1    16.411 310.55 71.372
## + x3    1     2.135 324.83 72.631
##
## Step:  AIC=50.78
## y ~ x8
##
##        Df Sum of Sq     RSS    AIC
## + x2    1    64.934  83.938 36.741
## + x5    1    11.607 137.265 50.512
## <none>              148.872 50.785
## + x1    1     6.636 142.236 51.508
## + x3    1     6.368 142.504 51.561
## + x4    1     6.345 142.527 51.565
## + x7    1     0.974 147.898 52.601
## + x6    1     0.487 148.385 52.693
## + x9    1     0.008 148.864 52.783
##
## Step:  AIC=36.74
## y ~ x8 + x2
##
##        Df Sum of Sq    RSS    AIC
## + x7    1   14.0682 69.870 33.604
## + x1    1   11.1905 72.748 34.734
## + x3    1    8.9010 75.037 35.602
## + x5    1    5.8147 78.124 36.730
## <none>              83.938 36.741
## + x9    1    2.0256 81.913 38.057
## + x6    1    1.3216 82.617 38.296
## + x4    1    0.0161 83.922 38.735
##
## Step:  AIC=33.6
## y ~ x8 + x2 + x7
##
##        Df Sum of Sq    RSS    AIC
## + x9    1    4.8657 65.004 33.583
## <none>              69.870 33.604
## + x3    1    1.3873 68.483 35.043
## + x4    1    0.9792 68.891 35.209
## + x1    1    0.9022 68.968 35.240
## + x6    1    0.4879 69.382 35.408
## + x5    1    0.2987 69.571 35.484
##
## Step:  AIC=33.58
## y ~ x8 + x2 + x7 + x9
##
##        Df Sum of Sq    RSS    AIC
## <none>              65.004 33.583
## + x1    1   1.86452 63.140 34.768
## + x4    1   1.74260 63.262 34.822
## + x3    1   0.70148 64.303 35.279
## + x6    1   0.45071 64.554 35.388
## + x5    1   0.32667 64.678 35.442
```

```
##
## Call:
## lm(formula = y ~ x8 + x2 + x7 + x9, data = data)
##
## Coefficients:
## (Intercept)           x8           x2           x7           x9
##   -1.821703    -0.004015     0.003819     0.216894    -0.001635
```

The model selected with forward selection is the model with predictors x2, x7, x8, and x9 as so:

y = -1.821703 + 0.003819x2 + 0.216894x7 - 0.004015x8 - 0.001635x9

(d) Run backward elimination, starting with the model with all predictors. Report the predictors and the estimated coefficients of the model selected.

```
step(regfull, scope=list(lower=reginter, upper=regfull), direction="backward")
```

```
## Start:  AIC=41.48
## y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9
##
##        Df Sum of Sq     RSS    AIC
## - x5    1     0.000  60.293 39.476
## - x1    1     0.549  60.842 39.730
## - x3    1     0.746  61.039 39.821
## - x6    1     0.803  61.096 39.847
## - x4    1     1.968  62.261 40.376
## - x7    1     3.451  63.744 41.035
## <none>               60.293 41.476
## - x9    1     5.348  65.642 41.856
## - x8    1    12.072  72.365 44.587
## - x2    1    62.448 122.741 59.380
##
## Step:  AIC=39.48
## y ~ x1 + x2 + x3 + x4 + x6 + x7 + x8 + x9
##
##        Df Sum of Sq     RSS    AIC
## - x1    1     0.553  60.846 37.732
## - x3    1     0.750  61.043 37.822
## - x6    1     0.818  61.111 37.854
## - x4    1     2.053  62.346 38.414
## - x7    1     3.859  64.152 39.213
## <none>               60.293 39.476
## - x9    1     5.351  65.644 39.857
## - x8    1    12.086  72.379 42.592
## - x2    1    66.979 127.272 58.395
##
## Step:  AIC=37.73
## y ~ x2 + x3 + x4 + x6 + x7 + x8 + x9
##
##        Df Sum of Sq     RSS    AIC
## - x6    1     0.690  61.536 36.048
## - x3    1     1.715  62.561 36.510
## - x4    1     3.051  63.897 37.102
## <none>               60.846 37.732
## - x9    1     4.852  65.698 37.880
## - x7    1     8.961  69.807 39.579
## - x8    1    16.599  77.445 42.486
## - x2    1    67.010 127.856 56.524
##
## Step:  AIC=36.05
## y ~ x2 + x3 + x4 + x7 + x8 + x9
##
##        Df Sum of Sq     RSS    AIC
## - x3    1     1.726  63.262 34.822
## - x4    1     2.767  64.303 35.279
## <none>               61.536 36.048
## - x9    1     4.831  66.367 36.164
## - x7    1     9.390  70.926 38.024
## - x8    1    18.314  79.851 41.343
## - x2    1    66.447 127.984 54.552
##
## Step:  AIC=34.82
## y ~ x2 + x4 + x7 + x8 + x9
##
##        Df Sum of Sq     RSS    AIC
## - x4    1     1.743  65.004 33.583
## <none>               63.262 34.822
## - x9    1     5.629  68.891 35.209
## - x8    1    17.701  80.962 39.730
## - x7    1    18.583  81.845 40.033
## - x2    1    75.598 138.860 54.835
##
## Step:  AIC=33.58
## y ~ x2 + x7 + x8 + x9
##
##        Df Sum of Sq     RSS    AIC
## <none>               65.004 33.583
## - x9    1     4.866  69.870 33.604
## - x7    1    16.908  81.913 38.057
## - x8    1    23.299  88.303 40.160
## - x2    1    82.892 147.897 54.601
```

```
##
## Call:
## lm(formula = y ~ x2 + x7 + x8 + x9, data = data)
##
## Coefficients:
## (Intercept)           x2           x7           x8           x9
##   -1.821703     0.003819     0.216894    -0.004015    -0.001635
```

The model selected with backward selection is the model with predictors x2, x7, x8, and x9 as so:

y = -1.821703 + 0.003819x2 + 0.216894x7 - 0.004015x8 - 0.001635x9

(e) The PRESS statistic can be used in model validation as well as a criteria for model selection. Unfortunately, the regsubsets() function from the leaps package does not compute the PRESS statistic.

Write a function that computes the PRESS statistic for a regression model. Hint: the diagonal elements from the hat matrix can be found using the lm.influence() function.

```
press = function(reg){
  x = reg$residuals/(1-lm.influence(reg)$hat)
  sum(x^2)
}
```

(f) Using the function you wrote in part 1e, calculate the PRESS statistic for your regression model with x2, x7, x8, x9 as predictors. Calculate the R2 Prediction for this model, and compare this value with its R2. What comments can you make about the likely predictive performance of this model?

```
1-(press(mod1)/sum(anova(mod1)$"Sum Sq"))
```

```
## [1] 0.7318984
```

```
summary(mod1)
```

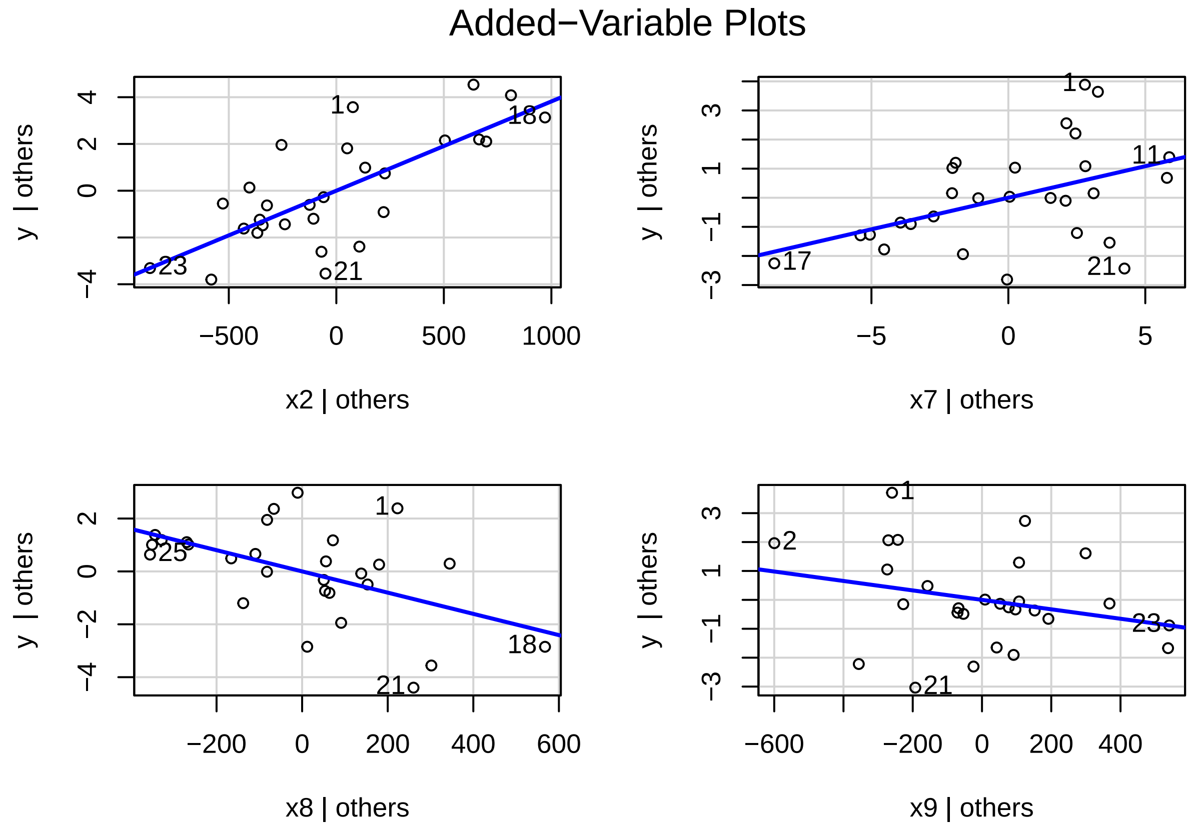
```
##
## Call:
## lm(formula = y ~ x2 + x7 + x8 + x9, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3519 -0.5612 -0.0856  0.6972  3.2802
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.8217034  7.7847061  -0.234  0.81705
## x2           0.0038186  0.0007051   5.416 1.67e-05 ***
## x7           0.2168941  0.0886759   2.446  0.02252 *
## x8          -0.0040149  0.0013983  -2.871  0.00863 **
## x9          -0.0016349  0.0012460  -1.312  0.20244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.681 on 23 degrees of freedom
## Multiple R-squared:  0.8012, Adjusted R-squared:  0.7666
## F-statistic: 23.17 on 4 and 23 DF,  p-value: 8.735e-08
```

The R sq Prediction for model 1 (x2, 7, 8, 9) is 0.7318984 and the multiple R sq value is 0.8012.

Looking at model 1, the R sq value is much higher than the R sq Prediction, meaning that model 1 is most likely to overfit the model (which we don't want).

(g) Create partial regression plots for this model. What are these plots telling us?

```
avPlots(mod1)
```



The partial regression plots can tell us how the model changes if we were to add the current predictor (assuming all other predictors are already in the model). So the slope of each regression line is the coefficient for each variable in the equation model. It can also suggest predictor variable transformations. For example, if we thought that x9 was more quadratic, we would need to add that higher order term to the model and test to see if there was significant change to the model.

(h) Using externally studentized residuals, do we have any outliers? What teams are these?

```
rstudent(mod1)
```

```
##            1            2            3            4            5            6 
##  2.212417494  0.698462761  1.980618602  1.436316628  0.015356592 -0.219873645 
##            7            8            9           10           11           12 
## -1.139124789  0.138984972  1.126646341 -1.531860339  0.079120509  1.020566939 
##           13           14           15           16           17           18 
## -0.031051813 -0.081865691 -1.870882721  0.377244734 -0.271115712 -0.430064184 
##           19           20           21           22           23           24 
##  0.305893844 -0.107708581 -2.309621577  0.943200983 -0.000315582 -0.347870613 
##           25           26           27           28 
## -0.530163595 -0.330278572 -0.085634289 -1.021220969
```

```
exsr = rstudent(mod1)
n = dim(data)[1]
p = 5
crit = qt(1-0.05/(2*n), n-1-p)
crit
```

```
## [1] 3.552167
```

```
exsr[abs(exsr)>crit]
```

```
## named numeric(0)
```

There are no detected outliers using the the crit value as a cut off for the studentized residuals.

(i) Do we have any high leverage data points for this multiple linear regression? What teams are these?

```
l = lm.influence(mod1)$hat
l[l>2*p/n]
```

```
##        18
## 0.3944162
```

An identified high leverage point is observation (team) 18.

(j) Use DFFITSi, DFBETASj,i, and Cook's distance to check for influential observations. What teams are influential?

```
## DFFITS
dff = dffits(mod1)
dff[abs(dff)>2*sqrt(p/n)]
```

```
## named numeric(0)
```

```
## DFBETA
dfb = dfbetas(mod1)
abs(dfb)>2/sqrt(n)
```

```
##    (Intercept)    x2    x7    x8    x9
## 1        FALSE FALSE FALSE  TRUE  TRUE
## 2        FALSE FALSE FALSE FALSE FALSE
## 3        FALSE FALSE FALSE FALSE FALSE
## 4        FALSE  TRUE FALSE FALSE FALSE
## 5        FALSE FALSE FALSE FALSE FALSE
## 6        FALSE FALSE FALSE FALSE FALSE
## 7        FALSE FALSE FALSE FALSE FALSE
## 8        FALSE FALSE FALSE FALSE FALSE
## 9        FALSE FALSE FALSE FALSE FALSE
## 10       FALSE FALSE FALSE  TRUE FALSE
## 11       FALSE FALSE FALSE FALSE FALSE
## 12       FALSE FALSE FALSE FALSE FALSE
## 13       FALSE FALSE FALSE FALSE FALSE
## 14       FALSE FALSE FALSE FALSE FALSE
## 15       FALSE FALSE FALSE FALSE  TRUE
## 16       FALSE FALSE FALSE FALSE FALSE
## 17       FALSE FALSE FALSE FALSE FALSE
## 18       FALSE FALSE FALSE FALSE FALSE
## 19       FALSE FALSE FALSE FALSE FALSE
## 20       FALSE FALSE FALSE FALSE FALSE
## 21        TRUE FALSE  TRUE  TRUE FALSE
## 22       FALSE FALSE FALSE FALSE FALSE
## 23       FALSE FALSE FALSE FALSE FALSE
## 24       FALSE FALSE FALSE FALSE FALSE
## 25       FALSE FALSE FALSE FALSE FALSE
## 26       FALSE FALSE FALSE FALSE FALSE
## 27       FALSE FALSE FALSE FALSE FALSE
## 28       FALSE FALSE FALSE FALSE FALSE
```

```
## Cooks D
cd = cooks.distance(mod1)
cd[cd>qf(0.5,p,n-p)]
```

```
## named numeric(0)
```

Observation 1 is influential for the associated coefficient for x8 and x9. Observation 4 is influential for the associated coefficient for x2. Observation 10 is influential for the associated coefficient for x8. Observation 15 is influential for the associated coefficient x9. Observation 21 is influential for the associated coefficient for the intercept, x7, and x8. This is going off of DFBETA. However, if we look at Cooks D and DFFITS cutoffs, there are no detected influential observations.

Question 2

```
Data<-faraway::wcgs
set.seed(6021) ##for reproducibility to get the same split
sample<-sample.int(nrow(Data), floor(.50*nrow(Data)), replace = F)
train<-Data[sample, ] ##training data frame
test<-Data[-sample, ] ##test data frame
```

(a) Before fitting a model, create some data visualizations to explore the relationship between these predictors and whether a middle-aged male develops coronary heart disease.

```
v1 = ggplot(train)+
  geom_density(aes(x=age, fill=chd), alpha=0.5)+
  scale_fill_manual(values=c("#38761d", "#e06666"))+
  theme_bw()
v2 = ggplot(train)+
  geom_density(aes(x=sdp, fill=chd), alpha=0.5)+
  scale_fill_manual(values=c("#38761d", "#e06666"))+
  theme_bw()
v3 = ggplot(train)+
  geom_density(aes(x=dbp, fill=chd), alpha=0.5)+
  scale_fill_manual(values=c("#38761d", "#e06666"))+
  theme_bw()
v4 = ggplot(train)+
  geom_density(aes(x=cigs, fill=chd), alpha=0.5)+
  scale_fill_manual(values=c("#38761d", "#e06666"))+
  theme_bw()
```

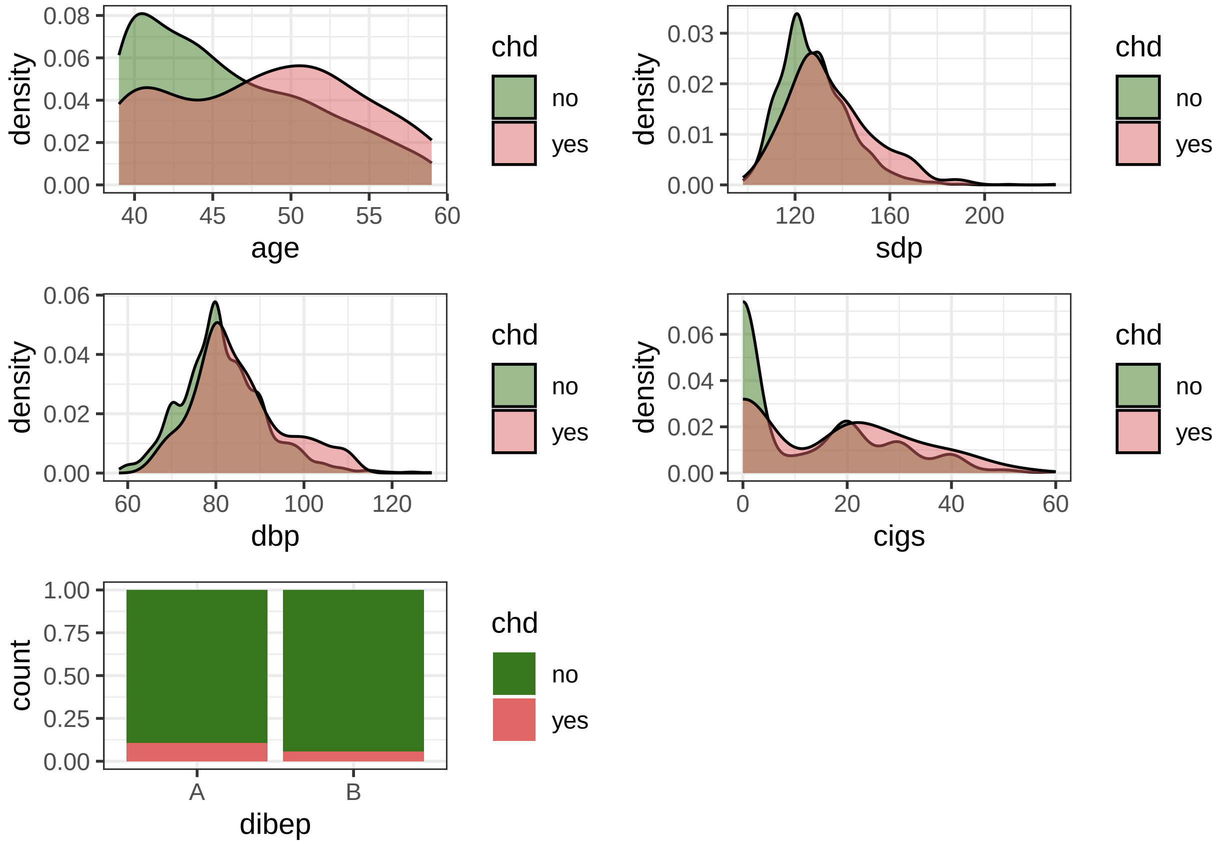
```
v5 = ggplot(train)+
  geom_bar(aes(x=dibep, fill=chd),  position = "fill")+
  scale_fill_manual(values=c("#38761d", "#e06666"))+
  theme_bw()
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
gridExtra::grid.arrange(v1, v2, v3, v4, v5, ncol = 2, nrow = 3)
```



(b) Use R to fit the logistic regression model using all the predictors listed above, and write the estimated logistic regression equation.

```
mod = glm(chd~age+sdp+dbp+cigs+dibep, data=train, family=binomial)
summary(mod)
```

```
##
## Call:
## glm(formula = chd ~ age + sdp + dbp + cigs + dibep, family = binomial,
##     data = train)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.308765   1.080141  -7.692 1.45e-14 ***
## age          0.060212   0.016604   3.626 0.000287 ***
## sdp          0.015119   0.008805   1.717 0.085950 .
## dbp          0.012026   0.014345   0.838 0.401818
## cigs         0.021366   0.006095   3.506 0.000456 ***
## dibepB      -0.526914   0.198429  -2.655 0.007921 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 893.04  on 1576  degrees of freedom
## Residual deviance: 837.55  on 1571  degrees of freedom
## AIC: 849.55
##
## Number of Fisher Scoring iterations: 5
```

chd = -8.308765 + 0.060212age + 0.015119sdp + 0.012026dbp + 0.021366cigs - 0.526914dibep, where dibep is 1 if passive behaviour type and 0 if aggressive behaviour type

(c) Interpret the estimated coefficient for cigs in context.

The log-odds of developing coronary heart disease increases by 0.021366 for every cigarettes smoked per day while keeping all other predictors constant.

(d) Interpret the estimated coefficient for dibep in context.

The log-odds of developing coronary heart disease decreases by 0.526914 for passive behaviour type people while keeping all other predictors constant.

(e) What are the estimated odds of developing heart disease for an adult male who is 45 years old, has a systolic blood pressure of 110 mm Hg, diastolic blood pressure of 70 mm Hg, does not smoke, and has type B personality? What is this person’s corresponding probability of developing heart disease?

```
newd = data.frame(age=45, sdp=110, dbp=70, cigs=0, dibep="B")
exp(predict(mod, newd))
```

```
##          1
## 0.02675027
```

The estimated odds of this person developing coronary heart disease is 0.02675027.

(f) Carry out the relevant hypothesis test to check if this logistic regression model with five predictors is useful in estimating the odds of heart disease. Clearly state the null and alternative hypotheses, test statistic, and conclusion in context.

```
gstat = mod$null.deviance-mod$deviance
gstat
```

```
## [1] 55.49501
```

```
qchisq(1-0.05, 5)
```

```
## [1] 11.0705
```

```
1-pchisq(gstat, 5)
```

```
## [1] 1.032455e-10
```

H0: beta(1, 2, 3, 4, 5)==0, the model is not adequate at estimating the odds of heart disease. HA: at least one beta(1,2,3,4,5)=/=0, at least one of the predictors is significant at estimating the odds of heart disease (the model is adequate). G-stat: 55.49501, chi-crit: 9.487729, pval: 1.032455e-10 ==> reject the H0 Conclusion: There is enough evidence to support the full model as at least one of these variables is adequate at estimating the odds of developing heart disease.

(g) Suppose a co-worker of yours suggests fitting a logistic regression model without the two blood pressure variables. Carry out the relevant hypothesis test to check if this model without the blood pressure variables should be chosen over the previous model with all five predictors.

```
mod2g = glm(chd~age+cigs+dibep, data=train, family=binomial)
gstat2 = mod2g$deviance-mod$deviance
gstat2
```

```
## [1] 13.70587
```

```
qchisq(1-0.05, 2)
```

```
## [1] 5.991465
```

```
1-pchisq(gstat2, 2)
```

```
## [1] 0.00105635
```

H0: beta(2,3)==0 HA: at least one=/=0 G-stat: 13.70587, chi-crit: 3.841459, pval: 0.00105635 ==> reject H0 Conclusion: There is enough evidence to support the full model and keep the two variables.

(h) Based on the Wald test, is diastolic blood pressure a significant predictor of heart disease, when the other predictors are already in the model?

The wald test suggests that diastolic blood pressure is not a significant predictor of heart disease.

(i) Based on all the analysis performed, which of these predictors would you use in your logistic regression model?

I would keep all of the variables in my model. Maybe I would consider dropping diastolic blood pressure too because it has a really high p value.

(j) Fit a logistic regression model based on your answer in part 2i. Based on the estimated coefficients of your logistic regression, briefly comment on the relationship between the predictors and the (log) odds of developing heart disease.

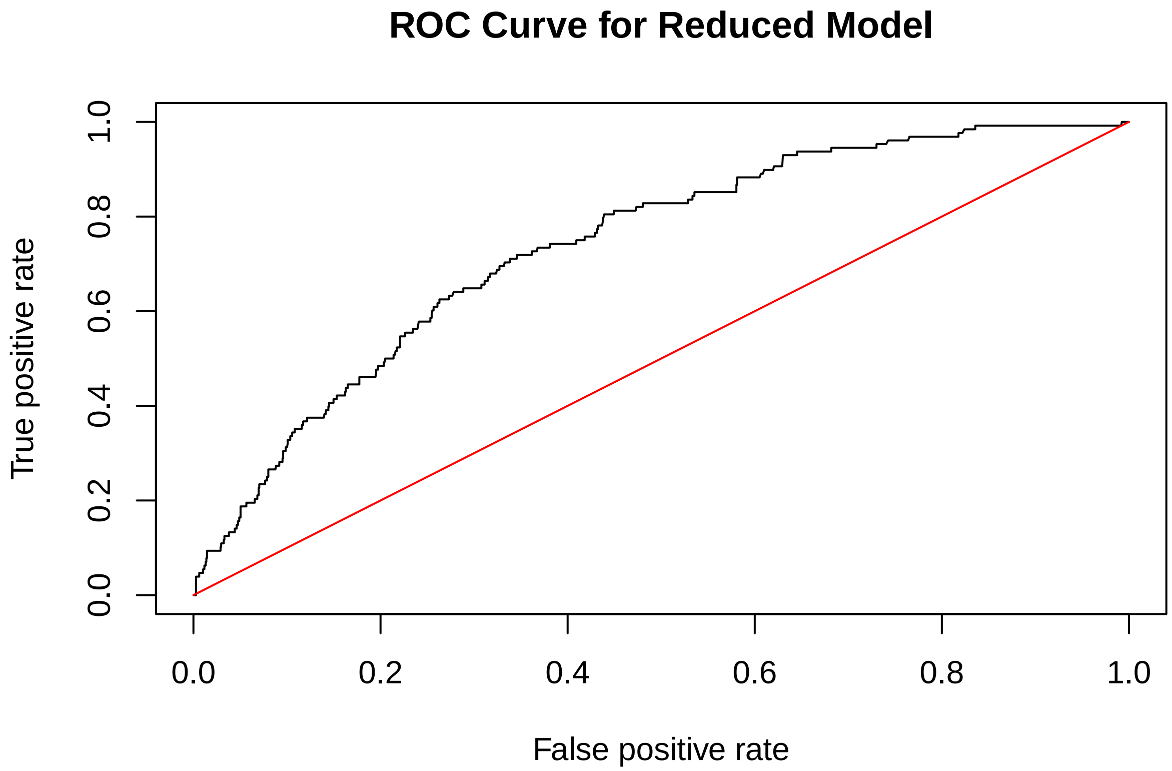
```
mod3g = glm(chd~age+sdp+cigs+dibep, data=train, family=binomial)
summary(mod3g)
```

```
##
## Call:
## glm(formula = chd ~ age + sdp + cigs + dibep, family = binomial,
##     data = train)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.065578   1.036178  -7.784 7.03e-15 ***
## age          0.060880   0.016560   3.676 0.000237 ***
## sdp          0.020757   0.005595   3.710 0.000207 ***
## cigs         0.020642   0.006035   3.421 0.000625 ***
## dibepB      -0.531792   0.198281  -2.682 0.007318 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 893.04  on 1576  degrees of freedom
## Residual deviance: 838.25  on 1572  degrees of freedom
## AIC: 848.25
##
## Number of Fisher Scoring iterations: 5
```

For an increase in age, sdp, or cig, the estimated odds of developing heart disease increases (keeping all other variables constant), but for people with a passive behaviour type, their estimated odds of developing heart disease decreases (keeping all other variables constant).

(k) Validate your logistic regression model using an ROC curve. What does your ROC curve tell you?

```
library(ROCR)
p=predict(mod3g, newdata=test, type="response")
r=ROCR::prediction(p, test$chd)
roc_result=ROCR::performance(r, measure="tpr", x.measure="fpr")
plot(roc_result, main="ROC Curve for Reduced Model")
lines(x = c(0,1), y = c(0,1), col="red")
```



Our ROC curve tells us how good our model is at estimating heart disease development. If our model was really good at correctly predicting heart disease development, then the curve would be closer to the upper left of the plot. In this case, the ROC curve suggests that our model is better at estimating heart disease development as opposed to just guessing randomly.

(l) Find the AUC associated with your ROC curve. What does your AUC tell you?

```
auc=performance(r, measure = "auc")
auc@y.values
```

```
## [[1]]
## [1] 0.7371679
```

The AUC is the area under the ROC curve for our regression model. In this case, our cutoff was at 0.5 and our AUC is 0.74, indicating that our model does a better job at estimating heart disease development as opposed to just guessing randomly.

(m) Create a confusion matrix using a cutoff of 0.5. Report the accuracy, true positive rate (TPR), and false positive rate (FPR) at this cutoff.

```
table(test$chd, p>0.5)
```

```
##
##       FALSE
##   no   1449
##   yes   128
```

```
n=dim(test)[1]
1449/n
```

```
## [1] 0.9188332
```

TPR: 0/128 = 0 FPR: 0/1449 = 0 Accuracy: 1449/(1449+128) = 0.9188332

(n) Based on the confusion matrix in part 2m, a classmate says the logistic regression at this cutoff is as good as a "no information classifier". Do you agree with your classmate's statement? Briefly explain.

Yes, because this is an unbalanced sample. Meaning at a threshold of 0.5, the model is just guessing at random. But most people (definitely not 50% of the population) don't have heart disease, so we would need to lower the threshold for this to make a little more sense.

(o) Discuss if the threshold should be adjusted. Will it be better to raise or lower the threshold? Briefly explain.

I would suggest lowering the threshold to 0.1 or 0.05 instead of 0.5 in order to get a little more information and to balance things out.

(p) Based on your answer in part 2o, adjust the threshold accordingly, and create the corresponding confusion matrix. Report the accuracy, TPR, and FPR for this threshold.

```
table(test$chd, p>0.05)
```

```
##
##       FALSE TRUE
##   no    521  928
##   yes     9  119
```

```
n=dim(test)[1]
```

(q) Comment on the results from the confusions matrices in parts 2m and 2p. What do you think is happening?

TPR and accuracy is sacrificed, but now the FPR

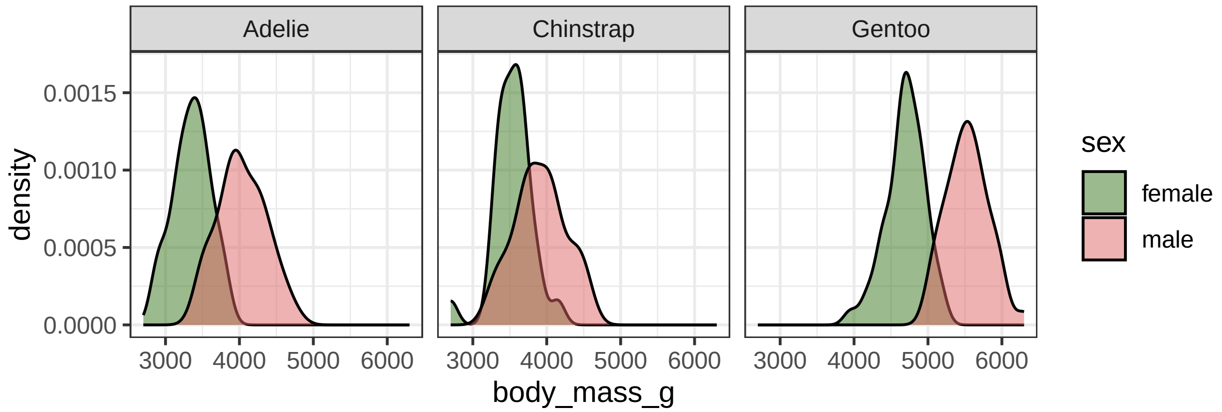
Question 3

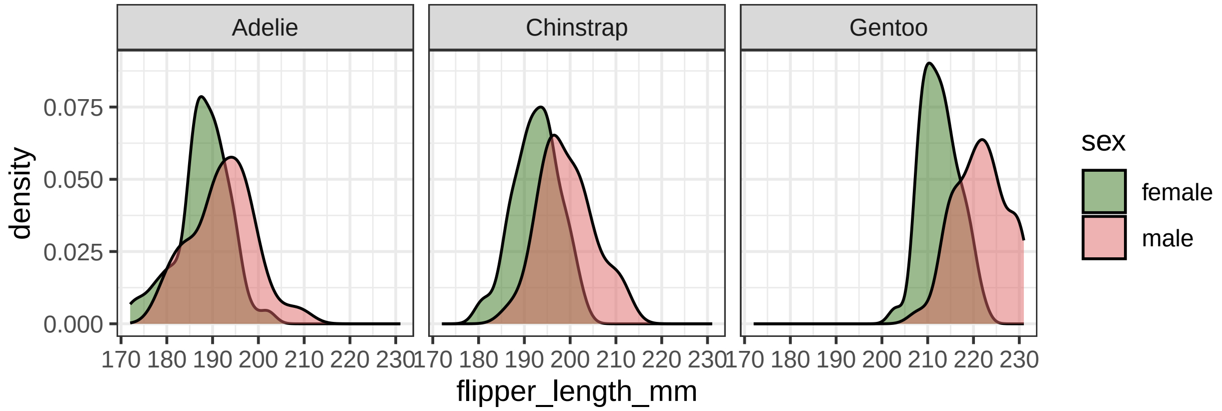
```
library(palmerpenguins)
Data<-palmerpenguins::penguins
##remove penguins with gender missing
Data<-Data[complete.cases(Data[ , 7]),-c(2,8)]
##80-20 split
set.seed(1)
sample<-sample.int(nrow(Data), floor(.80*nrow(Data)), replace = F)
train<-Data[sample, ]
test<-Data[-sample, ]
```

(a) Create some data visualizations to explore the relationship between the various body measurements and the gender of penguins. Be sure to briefly comment on your data visualizations.

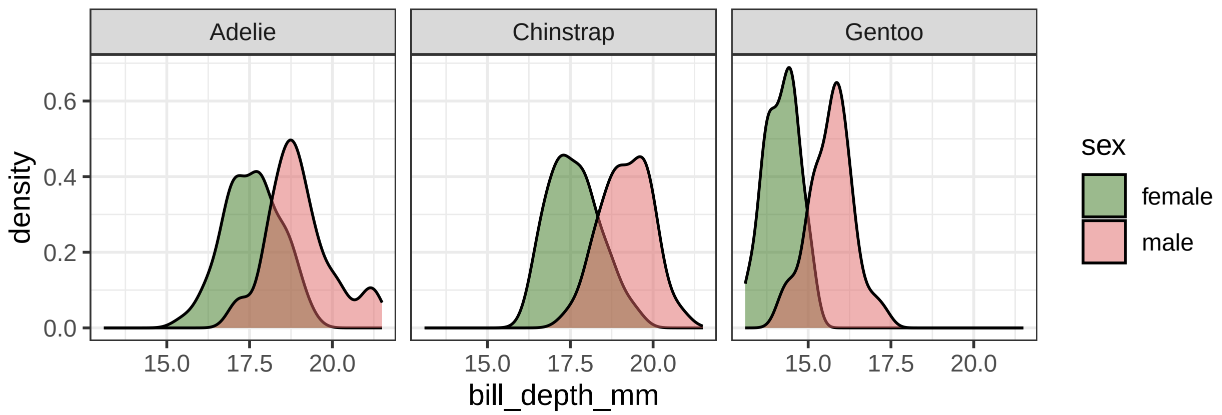
```
v1 = ggplot(train)+
  geom_density(aes(x=body_mass_g, fill=sex), alpha=0.5)+
  scale_fill_manual(values=c("#38761d", "#e06666"))+
  facet_wrap(~species)+
  theme_bw()
v2 = ggplot(train)+
  geom_density(aes(x=flipper_length_mm, fill=sex), alpha=0.5)+
  scale_fill_manual(values=c("#38761d", "#e06666"))+
  facet_wrap(~species)+
  theme_bw()
v3 = ggplot(train)+
  geom_density(aes(x=bill_depth_mm, fill=sex), alpha=0.5)+
  scale_fill_manual(values=c("#38761d", "#e06666"))+
  facet_wrap(~species)+
  theme_bw()
v4 = ggplot(train)+
  geom_density(aes(x=bill_length_mm, fill=sex), alpha=0.5)+
  scale_fill_manual(values=c("#38761d", "#e06666"))+
  facet_wrap(~species)+
  theme_bw()
v5 = ggplot(train)+
  geom_bar(aes(x=species, fill=sex),  position = "fill")+
  scale_fill_manual(values=c("#38761d", "#e06666"))+
  theme_bw()
```

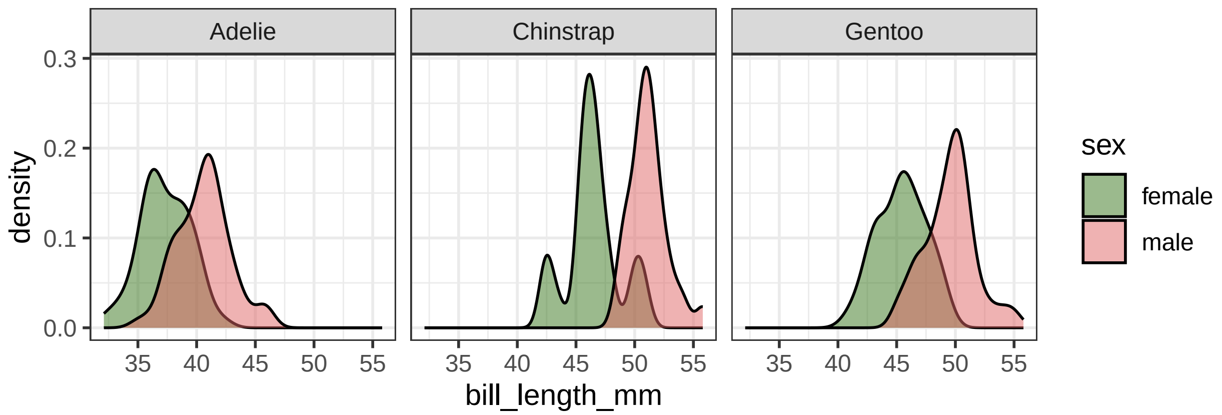
```
gridExtra::grid.arrange(v1, v2, ncol = 1)
```



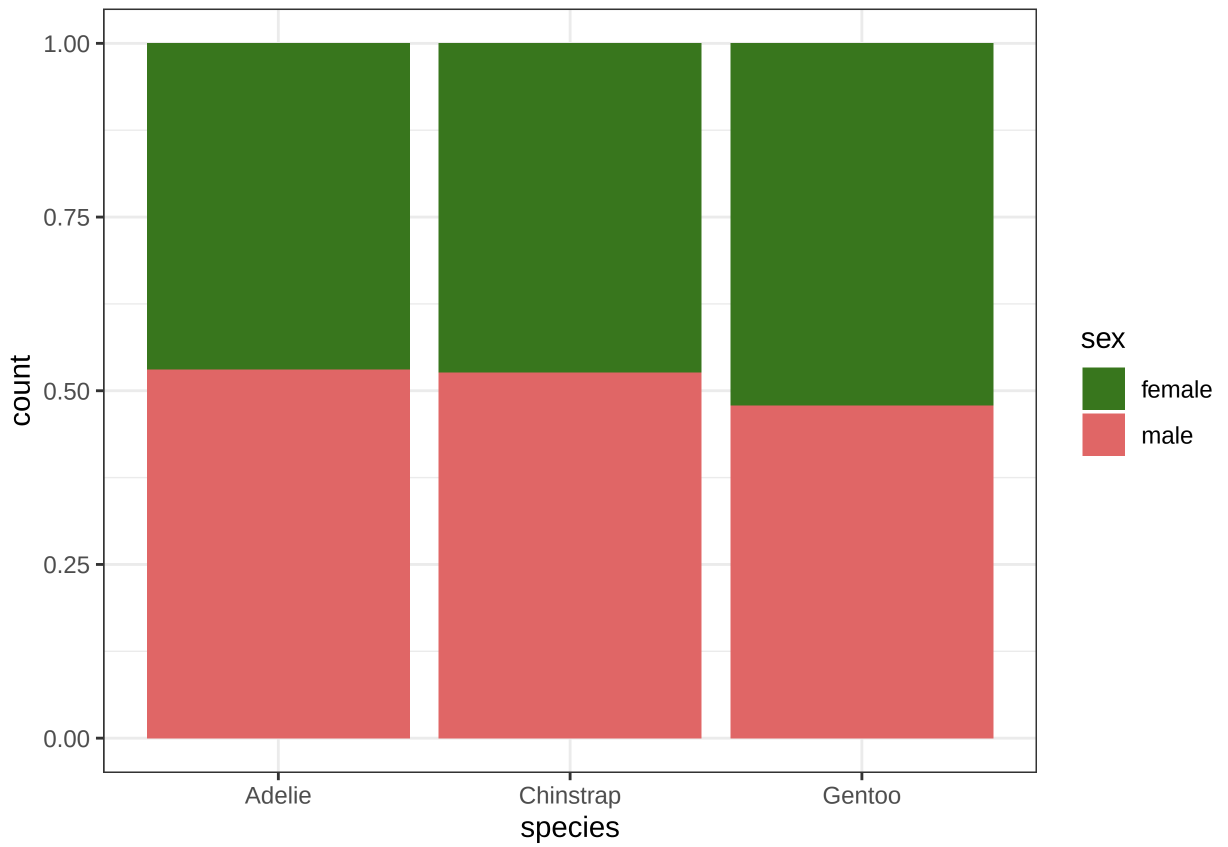


```
gridExtra::grid.arrange(v3, v4, ncol = 1)
```


Adelie
Chinstrap
Gentoo
density
0.0
0.2
0.4
0.6
15.0
17.5
20.0
bill_depth_mm
sex
🟩 female
🟥 male


Adelie
Chinstrap
Gentoo
density
0.0
0.1
0.2
0.3
35
40
45
50
55
bill_length_mm
sex
🟩 female
🟥 male

```
v5
```



While controlling for the penguin species, there appears to be a difference in body measurements between the sexes. Males seem to generally have larger/bigger body parts than their female counterparts. The difference between body measurements may or may not be significant though. The proportion of male and female penguins also seem to be balanced within each species.

(b) Use R to fit the logistic regression model. Based on the results of the Wald tests for the individual coefficients, which predictor(s) appears to be insignificant in the model?

```
mod3full = glm(sex~., family="binomial", data=train)
summary(mod3full)
```

```
##
## Call:
## glm(formula = sex ~ ., family = "binomial", data = train)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -94.355394  17.638204  -5.349 8.82e-08 ***
## speciesChinstrap  -10.608813   2.634752  -4.026 5.66e-05 ***
## speciesGentoo     -10.384568   3.565641  -2.912  0.00359 **
## bill_length_mm      1.025200   0.238593   4.297 1.73e-05 ***
## bill_depth_mm       2.287977   0.516595   4.429 9.47e-06 ***
## flipper_length_mm  -0.088318   0.065040  -1.358  0.17450
## body_mass_g         0.008094   0.001662   4.871 1.11e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 368.619  on 265  degrees of freedom
## Residual deviance:  68.297  on 259  degrees of freedom
## AIC: 82.297
##
## Number of Fisher Scoring iterations: 8
```

Flipper length appears to be insignificant according to the z test.

(c) Based on your answer in part 3b, drop the predictor(s) and refit the logistic regression. Write out the estimated logistic regression equation. If you did not drop any predictor, write out the logistic regression equation from part 3b.

```r
mod3red = glm(sex~species+bill_length_mm+bill_depth_mm+body_mass_g, family="binomial", data=train)
summary(mod3red)
```

```
##
## Call:
## glm(formula = sex ~ species + bill_length_mm + bill_depth_mm +
##     body_mass_g, family = "binomial", data = train)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.032e+02  1.706e+01  -6.051 1.44e-09 ***
## speciesChinstrap -1.042e+01  2.544e+00  -4.096 4.20e-05 ***
## speciesGentoo    -1.238e+01  3.383e+00  -3.661 0.000251 ***
## bill_length_mm    9.513e-01  2.210e-01   4.303 1.68e-05 ***
## bill_depth_mm     2.099e+00  4.684e-01   4.481 7.41e-06 ***
## body_mass_g       7.714e-03  1.625e-03   4.746 2.07e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 368.619  on 265  degrees of freedom
## Residual deviance:  70.172  on 260  degrees of freedom
## AIC: 82.172
##
## Number of Fisher Scoring iterations: 8
```

The estimated logistics regression equation is as follows:

sex = -1.032e+02 - 1.042e+01Chin - 1.238e+01Gentoo + 9.513e-01BillLength + 2.099BillDepth + 7.714e-03BodyMass

(d) Based on your estimated logistic regression equation in part 3c, how would you generalize the relationship between some of the body measurement predictors and the (log) odds of a penguin being male?

For Chipstrap or Gentoo penguins, the log odds of them being male decreases, (while holding all other variables constant). For Adelie penguins, or penguins with larger bill lengths, bill depths, or body mass, the log odds of them being male increases (while holding all other variables constant).

(e) Based on your estimated logistic regression equation in part 3c, interpret the estimated coefficient for bill length contextually.

For each mm increase in bill length, the estimated log odds of a penguin being male increases by 0.9513 while holding all other variables constant.

(f) Consider a Gentoo penguin with bill length of 49mm, bill depth of 15mm, flipper length of 220mm, and body mass of 5700g. Based on your estimated logistic regression equation in part 3c, what are the log odds, odds, and probability that this penguin is male?

```
new3d = data.frame(species="Gentoo", bill_length_mm=49, bill_depth_mm=15, flipper_length_mm=220, body_ma
logodds = predict(mod3full, new3d)
odds = exp(logodds)
prob = odds/(1+odds)
logodds
```

```
##        1
## 6.519736
```

```
odds
```

```
##        1
## 678.3991
```

```
prob
```

```
##         1
## 0.9985281
```

For a Gentoo penguin with these measurements, the log odds of them being male is 6.519, the odds of them being male is 678.399, and the probability of them being male is 0.998.

(g) Conduct a relevant hypothesis test to assess if the logistic regression in part 3c is a useful model. Be sure to write out the null and alternative hypotheses, report the value of the test statistic, and write a relevant conclusion.

```
g3g = mod3red$null.deviance-mod3red$deviance
g3g
```

```
## [1] 298.4472
```

```
qchisq(1-0.05, 5)
```

```
## [1] 11.0705
```

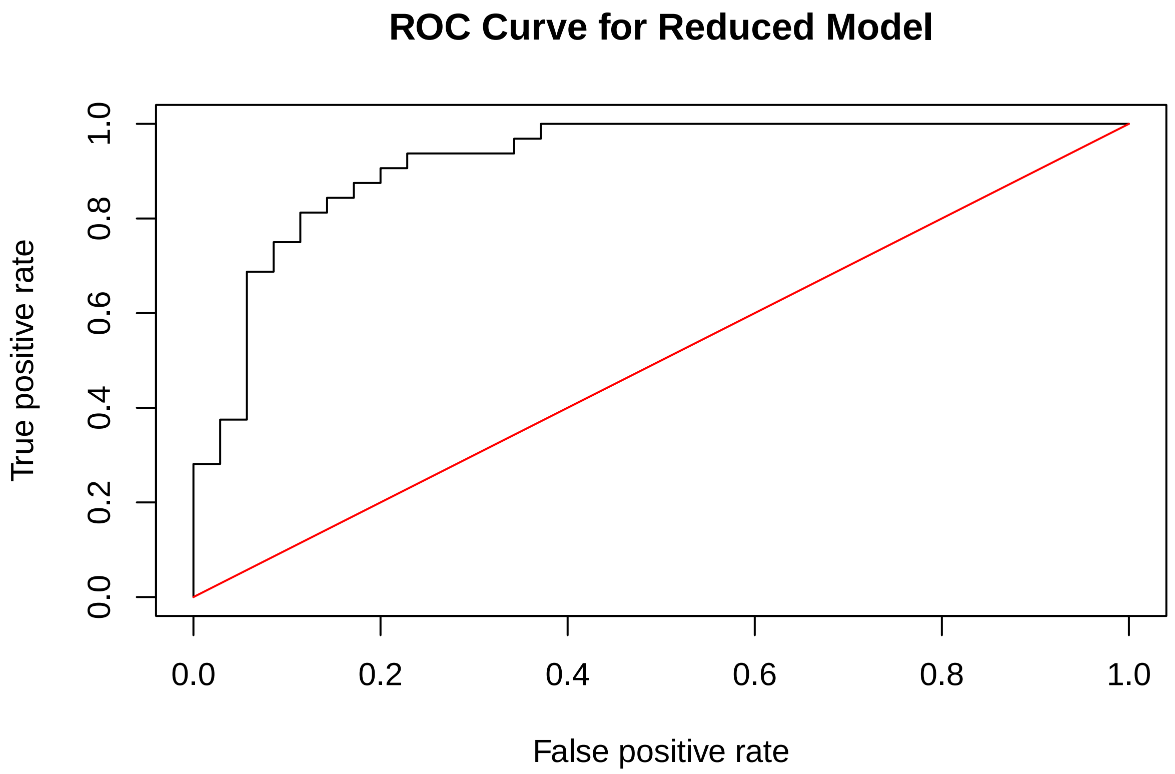
```
1-pchisq(g3g, 5)
```

```
## [1] 0
```

H0: beta(1,2,3,4,6)==0 (the model is not useful over an intercept model) HA: at least one=/=0 (the model is useful over an intercept model) G-stat: 298.4472, chi-crit: 11.0705, p-val: 0 ==> reject the H0 Conclusion: We have enough evidence that suggests that the current model is useful and should be chosen over an intercept-only model.

(h) Validate your model from part 3c on the test data by creating an ROC curve. What does your ROC curve tell you?

```
p3=predict(mod3red, newdata=test, type="response")
r3=ROCR::prediction(p3, test$sex)
roc_result=ROCR::performance(r3, measure="tpr", x.measure="fpr")
plot(roc_result, main="ROC Curve for Reduced Model")
lines(x = c(0,1), y = c(0,1), col="red")
```



Our ROC curve tells us how good our model is at estimating the sex of penguins. If our model was really good at correctly predicting the sex, then the curve would be closer to the upper left of the plot. In this case, it is considerably close to the upper left corner (noice).

(i) Find the AUC associated with your ROC curve. What does your AUC tell you?

```
auc3=ROCR::performance(r3, measure = "auc")
auc3@y.values
```

```
## [[1]]
## [1] 0.9214286
```

The AUC is the area under the ROC curve for our regression model. In this case, our cutoff was at 0.5 and our AUC is 0.9214286, indicating that our model does a better job at predicting penguin sex as opposed to just guessing randomly.

(j) Create a confusion matrix using a threshold of 0.5. What is the false positive rate? What is the false negative rate? What is error rate?

```
table(test$sex, p3>0.5)
```

```
##
##          FALSE TRUE
##   female    28    7
##   male       4   28
```

```
n=dim(test)[1]
n
```

```
## [1] 67
```

FPR: 7/(7+28) = 0.2 FNR: 4/(28+4) = 0.125 Error Rate: (7+4)/67 = 0.1641791

(k) Discuss if the threshold should be changed. If it should be changed, explain why, and create another confusion matrix with a different threshold.

No, I don't think we need to change the threshold. Based on the proportion bar chart, the proportion of male and female penguins for each species is around 0.5, so we should have our cut off be around the same.

Question 4

(a) The output below is obtained after using the step() function using forward selection, starting with a model with just the intercept term. What predictors are selected based on forward selection?

The selected predictors are discount, promo, and price.

(b) Your client asks you to explain what each step in the output shown above means. Explain the forward selection procedure to your client, for this output.

This forward selection uses a measurement called AIC. We want a model with a smaller AIC. A forward selection takes an intercept model (a model with no predictors), and tests the AIC for all possible 1-predictor models. It compares the AIC of these 1-predictor models with the intercept only model and chooses the model with the smallest AIC. In this case, it was the model with the discount predictor. Next, the selection takes that one predictor and tests the remaining predictors one at a time to see which 2-predictor model has the smallest AIC. So in this case, discount+promo had the smallest AIC. This goes on (testing 3-predictors, 4-predictors etc) until the selection can no longer find a model with a smaller AIC than the current model. The selected ended at discount+promo+price because adding time or adding nielsen to the model increased the AIC.

(c) Your client asks if he should go ahead and use the models selected in part 4a. What advice do you have for your client?

I would advise him to the model as a baseline. Testing for higher order terms and interactions are just as important and should be considered before going ahead and using the current model.

Question 5

(a) Calculate the externally studentized residual, ti, for observation 6. Will this be considered outlying?

```
SSres5 = 17*(40.13^2)
SSres5
```

```
## [1] 27377.09
```

```
h66 = 0.23960510
e6 = 120.829070
t6 = e6*(16/(SSres5*(1-h66)-e6^2))^0.5
t6
```

```
## [1] 6.129363
```

```
qt(1-(0.05/38), 16)
```

```
## [1] 3.556242
```

Since the externally studentized residual is greater than the crital value (magnitude wise), observation 6 can be considered an outlier.

(b) What is the leverage for observation 6? Based on the criterion that leverages greater than 2p/n are considered outlying in the predictor(s), is this observation high leverage?

```
4/19
```

```
## [1] 0.2105263
```

The leverage is 0.23960510 which is greater than the cut off of 0.21, indicating that this observation is high in leverage.

(c) Calculate the DFFITS for observation 6.

```
t6*(h66/(1-h66))^0.5
```

```
## [1] 3.440676
```

The DFFITS for obs 6 is 3.440676.

(d) Calculate Cook's distance for observation 6.

```
p1 = e6^2/(2*40.13^2)
p2 = h66/((1-h66)^2)
p1*p2
```

```
## [1] 1.878418
```

Cook's D for obs 6 is 1.878418.

(e) Would you say that observation 6 is influential, based on DFFITS and Cook's distance?

```
qf(0.5,2,17)
```

```
## [1] 0.7221933
```

```
2*sqrt(2/19)
```

```
## [1] 0.6488857
```

By comparing DFFITS and Cook's D for obs 6, where Cook's D of 1.89 is greater than it's cutoff of 0.7221933 and DFFITS of 3.44 is greater than it's cut off of 0.6488857, I would say that obs 6 is influential.

(f) Briefly describe the difference in what DFFITS and Cook's distance are measuring.

DIFFTS is a measurement between the value of an observation and the predicted value of that observation. So it's focused on how the model differs in terms of that one observation. Cook's Distance takes that to a broader sense where it focuses on how all of the observations are affected if one particular observation is gone.

TLDR; DFFITS = impact of predicted value for observation i if observation i is removed. Cook's D = impact of all predict3ed values (or coefs) if observation i is removed.