BIOST 546
WINTER QUARTER 2020

Homework # 2
Due Via Online Submission to Canvas: Tues, Feb 11 at 12 PM (Noon)

Instructions:

You may discuss the homework problems in small groups, but you must write up the final solutions and code yourself. Please turn in your code for the problems that involve coding. However, code without written answers will receive no credit. To receive credit, you must explain your answers and show your work. All plots should be appropriately labeled and legible, with axis labels, legends, etc., as needed.

Please remember — the easier you make it for the TA to find your answer, the easier it will be for him to give you credit for the problem!

1. Consider classification with K classes and one feature, i.e. $p = 1$.

   In lecture, we went through a detailed argument to see that the discriminant function for linear discriminant analysis (which assumes that an observation in the kth class is drawn from a $N(\mu_k, \sigma^2)$ distribution) is of the form given in Equation 4.13 of the textbook. **5+3 = 8 points**

   (a) Now consider quadratic discriminant analysis, which assumes that an observation in the kth class is drawn from a $N(\mu_k, \sigma_k^2)$ distribution. Using an argument similar to the one in class, derive the discriminant function for quadratic discriminant analysis. It should be similar, but not identical, to Equation 4.13. It should also look similar to Equation 4.23 (but not identical — Equation 4.23 is a little bit more complicated since it has $p > 1$).

   The QDA assumption states that for class k,

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$$

   . Recall in lecture that this implies

$$p(y = k|x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2)}{\sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2)}$$

.

Moreover, the classification rule of QDA is to classify a class k that maximizes $p(y = k|x)$, which is equivalent to maximizing

$$\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2)$$

.

Now since log is a monotonically increasing function (i.e., $a \geq b \implies \log a \geq \log b$), it's equivalent to maximizing

$$\log \pi_k - \log \sigma_k - \frac{1}{2\sigma_k^2}(x - \mu_k)^2$$

.

Note that we dropped $\log\sqrt{2\pi}$ since it's a constant. Now expand the last square term gives us

$$\delta_k(x) = -\frac{x^2}{2\sigma_k^2} + \frac{x \cdot \mu_k}{\sigma_k^2} - \frac{\mu_k^2}{2\sigma_k^2} + \log \pi_k - \log \sigma_k$$

.

(b) Comment on the difference between Equation 4.13 and your answer in (a). Explain how we can see that the discriminant functions are for linear discriminant analysis and quadratic discriminant analysis are *linear* and *quadratic*, respectively.

The decision function in Eqaution 4.13 is

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k$$

, which does not have the additional quadratic term compared to our answer in part (a). In particular, the decision function for QDA has the quadratic term $-\frac{x^2}{2\sigma_k^2}$ which makes $\delta_k(x)$ a *quadratic* funciton of x, where in the LDA case $\delta_k(x)$ is a *linear* function of x.

2. Choose a data set with $p = 2$ features X_1 and X_2, a qualitative response Y with $K = 3$ classes, and at least 15 observations per class. (If you have a data set with more than two features or more than three classes, then feel free to just select a subset of the features and classes so that you can use the data for this problem.) We are going to predict Y using X_1 and X_2. **3+3+3+3+3 = 15 points**

   (a) Briefly describe the data. Where did you get it? Describe the K classes and the p features. Explain the classification task in words (e.g. a sentence along the lines of "I will use the expression levels of genes ABC and DEF to predict whether a patient belongs to class G, H, or I.")

We will be revisiting Fisher's or Anderson's iris data set in `R` as promised!! This data set has 150 measurements $n = 150$ and 4 features ($p = 4$: sepal length, sepal width, petal length, and petal width). There are 3 classes and they correspond to 3 species of iris (Iris setosa, versicolor, and virginica).

(b) Fit an LDA model to the data. Make a plot with X_1 and X_2 on the horizontal and vertical axes, and with the observations displayed and colored according to their true class labels. On the plot, indicate which observations are incorrectly classified.

```
#### Q2 iris prediction problem
library(class)
data(iris)
# y: Species (setosa,versicolor,virginica)
# x1: Sepal.Length
# x2: Petal.Length
my_x1 <- iris$Sepal.Length
my_x2 <- iris$Petal.Length
my_y <- as.factor(iris$Species)

lda_model <- lda(Species ~ Sepal.Length+Petal.Length, data = iris)
lda_prediction <- predict(lda_model, iris)$class
lda_correct <- (lda_prediction==iris$Species)
pch_lda <- rep(16, times=length(lda_correct))
pch_lda[!lda_correct] <- 17

plot(x=my_x1,y=my_x2,
     col=my_y,
     pch=pch_lda,
     xlab = 'Sepal Length',
     ylab='Petal Length',
     main='Visualizing the prediction from LDA')
legend(7,3,legend=c("Setosa", "Versicolor", "Virginica"),
       col = c('black','red','green'),cex=1,pch =16)
legend(7,4,legend=c("Correct", "Incorrect"),cex=1,pch =c(16,17))
```

(c) Fit a QDA model to the data. Make a plot with X_1 and X_2 on the horizontal and vertical axes, and with the observations displayed and colored according to their true class labels. On the plot, indicate which observations are incorrectly classified.

```
qda_model <- qda(Species ~ Sepal.Length+Petal.Length, data = iris)
qda_prediction <- predict(qda_model, iris)$class
qda_correct <- (qda_prediction==iris$Species)
```

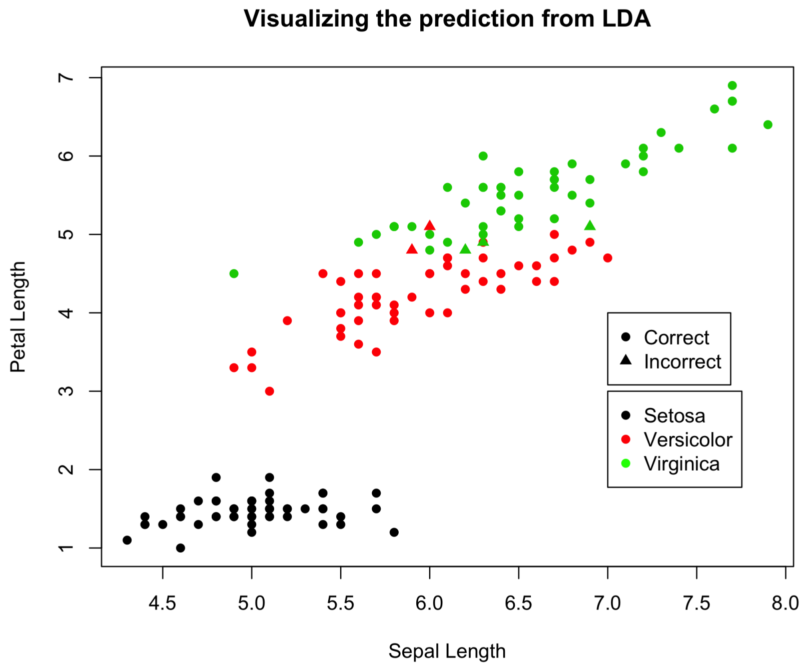

Figure 1: Plotting the predictions for LDA where we color the observations using their true class and plot the incorrect predictions using triangles instead of circles.

```
pch_qda <- rep(16, times=length(qda_correct))
pch_qda[!qda_correct] <- 17

plot(x=my_x1,y=my_x2,
     col=my_y,
     pch=pch_qda,
     xlab = 'Sepal Length',
     ylab='Petal Length',
     main='Visualizing the prediction from QDA')
legend(7,3,legend=c("Setosa", "Versicolor", "Virginica"),
       col = c('black','red','green'),cex=1,pch =16)
legend(7,4,legend=c("Correct", "Incorrect"),cex=1,pch =c(16,17))
```

(d) Fit a logistic regression model to the data. Make a plot with X_1 and X_2 on the horizontal and vertical axes, and with the observations displayed and colored according to their true class labels. On the plot, indicate which observations are incorrectly classified. **NOT graded!**

```
library(nnet)
multi_logistic <- multinom(Species ~ Sepal.Length+Petal.Length, data = iris,
pred_logistic <- predict(multi_logistic, newdata = iris)
```

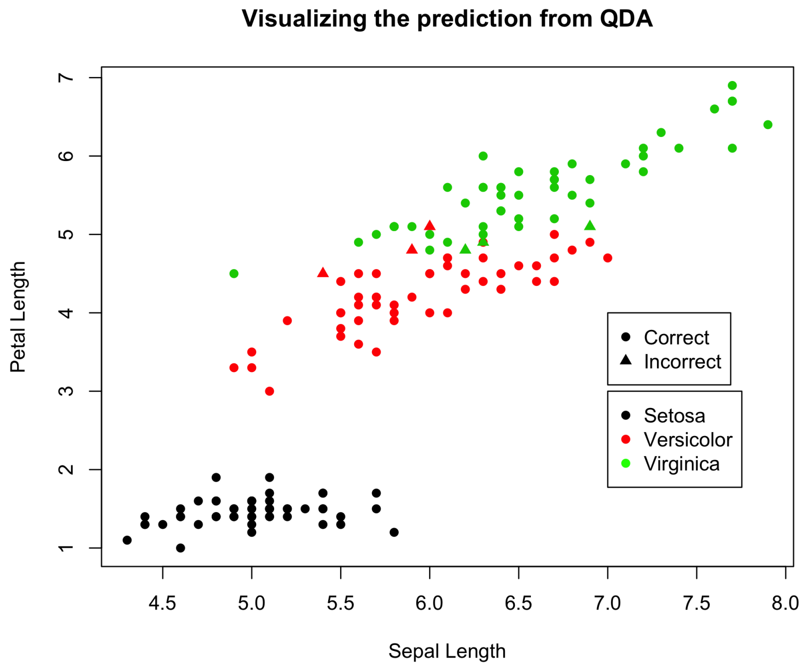

Figure 2: Plotting the predictions for QDA where we color the observations using their true class and plot the incorrect predictions using triangles instead of circles.

```
logistic_correct <- (pred_logistic==iris$Species)
pch_logistic <- rep(16, times=length(logistic_correct))
pch_logistic[!logistic_correct] <- 17

plot(x=my_x1,y=my_x2,
     col=my_y,
     pch=pch_logistic,
     xlab = 'Sepal Length',
     ylab='Petal Length',
     main='Visualizing the prediction from Logistic')
legend(7,3,legend=c("Setosa", "Versicolor", "Virginica"),
        col = c('black','red','green'),cex=1,pch =16)
legend(7,4,legend=c("Correct", "Incorrect"),cex=1,pch =c(16,17))
```

(e) Out of the three models, which one gave you the smallest training error? How does this relate to the bias-variance trade-off?

	Training error
LDA	0.033
QDA	0.040
Multiclass logistic	0.033

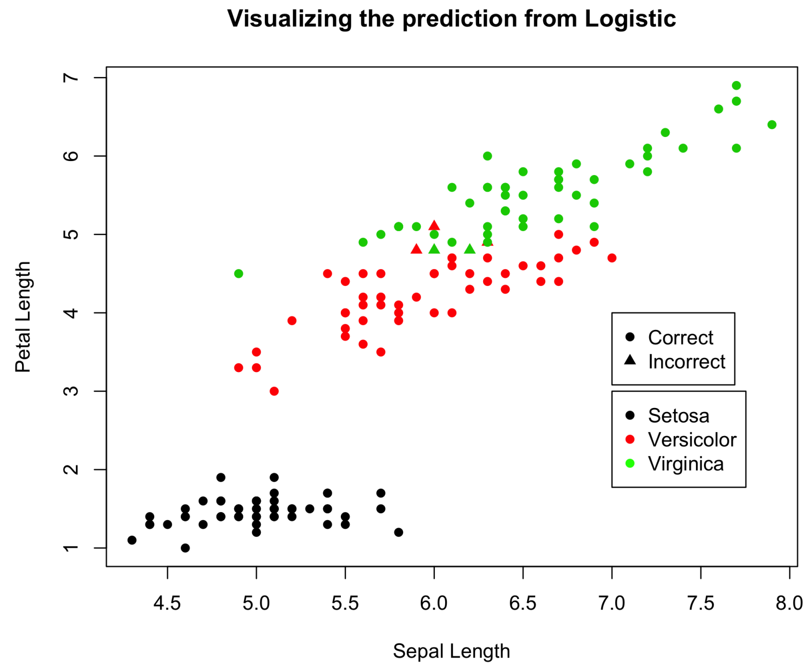

Figure 3: Plotting the predictions for multiclass logistic regression where we color the observations using their true class and plot the incorrect predictions using triangles instead of circles.

We note that the training error is very similar for all three models and QDA has very slightly higher training error (1 additional misclassification). This is in fact not what we would expect: since QDA is a more flexible model compared to LDA, we would expect it to have smaller training error in most cases. However in the particular example a linear decision boundary really fits the data well so all three models give us indistinguishable error rates. In principle, a more flexible model should almost always lead to a smaller training error.

(f) Which of these three models do you expect will give you the smallest test error? Explain your answer. How does this relate to the bias-variance trade-off?

We would expect the test error for LDA (or multiclass logistic regression) to be the smallest on this particular data set. Our reasoning is that since the training errors for LDA and QDA are very similar, the underlying decision boundary is approximately linear and the additional flexibility of QDA most likely will lead to overfitting.

3. Suppose we have a quantitative response Y, and two quantitative features X_1 and X_2. Let RSS_1 denote the residual sum of squares that results from fitting

the model

$$Y = \beta_0 + \beta_1 X_1 + \epsilon \tag{1}$$

using least squares. Let RSS_{12} denote the residual sum of squares that results from fitting the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \tag{2}$$

using least squares.

Perform the following procedure a whole lot of times (you will need to write a for loop to do this):

- Simulate Y, X_1, and X_2 with $n = 200$. You can generate each element of X_1 and X_2 independently from a $N(0, 1)$ distribution, and you can generate Y according to $Y = 3 + 2X_1 - X_2 + \epsilon$, where the elements of ϵ are independent draws from a $N(0, 1)$ distribution.
- Fit the models (1) and (2) using least squares.
- Compare the values of RSS_{12} and RSS_1.
- Compare the R^2 value for (1) to the R^2 value for (2).

Describe your findings. Which of the two models is more flexible? Which model has smaller training RSS, and which model has larger training R^2? How would you expect the two models to perform on test data? How do your findings relate to the bias-variance trade-off? **2+2+2+2 = 8 points, Extra credits: 5 points**

We ran the simulation 1,000 times using the following code:

```
   ### q3 simulation
set.seed(12345) # make your simulation reproducible!
sim_times <- 1000
rss_1_vec <- rep(NA,sim_times)
rss_12_vec <- rep(NA,sim_times)
r2_1_vec <- rep(NA,sim_times)
r2_2_vec <- rep(NA,sim_times)

for (i in 1:sim_times){
  x1 <- rnorm(mean=0, sd=1, n=200)
  x2 <- rnorm(mean=0, sd=1, n=200)
  error <- rnorm(mean=0, sd=1, n=200)
  y <- 3+2*x1-x2+error
  model_1 <- summary(lm(y~x1))
  model_2 <- summary(lm(y~x1+x2))
  # extract info from the fit lm model
  rss_1 <- sum((model_1$residuals)^2)
  rss_12 <- sum((model_2$residuals)^2)
```

```
  r2_1 <- model_1$r.squared
  r2_12 <- model_2$r.squared
  # store the results for current simulation
  rss_1_vec[i] <- rss_1
  rss_12_vec[i] <- rss_12
  r2_1_vec[i] <- r2_1
  r2_2_vec[i] <- r2_12
}

rb <- boxplot(decrease ~ treatment, data = OrchardSprays, col = "bisque")
title("Comparing boxplot()s and non-robust mean +/- SD")

library(latex2exp)
plot_q3 <- data.frame(RSS=c(rss_1_vec,rss_12_vec), R2=c(r2_1_vec,r2_2_vec),
           model = rep(c('Model (1)','Model (2)'),each = sim_times))

boxplot(RSS ~ model,
        data = plot_q3,
        col = "lightgray",
        ylab = 'RSS',
        main = TeX('Comparing $RSS_1$ and $RSS_{12}$'))

boxplot(R2 ~ model,
        data = plot_q3,
        col = "lightgray",
        ylab= TeX('$R^2$'),
        main = TeX('Comparing $R^2_1$ and $R^2_{12}$'))
```

The boxplots below show that RSS_1 is consistently larger than RSS_{12} and model (1)'s R^2 is consistently smaller than that of model (2). In other words, the more flexible model 2 has smaller training RSS and larger training R^2. We would expect the same observation on test data: We know compared to model 1, model 2 has smaller bias and larger variance. However since the predictor set in model 2 is *correctly specified*, it lands itself on the "sweet spot" of the bias-variance trade-off!

*****Extra Credit:***** Prove that $RSS_{12} \leq RSS_1$.

Using the definitions of least square estimates $\hat{\beta}$ and RSS lets us rewrite RSS_{12} as:

$$RSS_{12} = \min_{\beta_0,\beta_1,\beta_2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 \cdot X_{i1} - \beta_2 \cdot X_{i2})^2$$

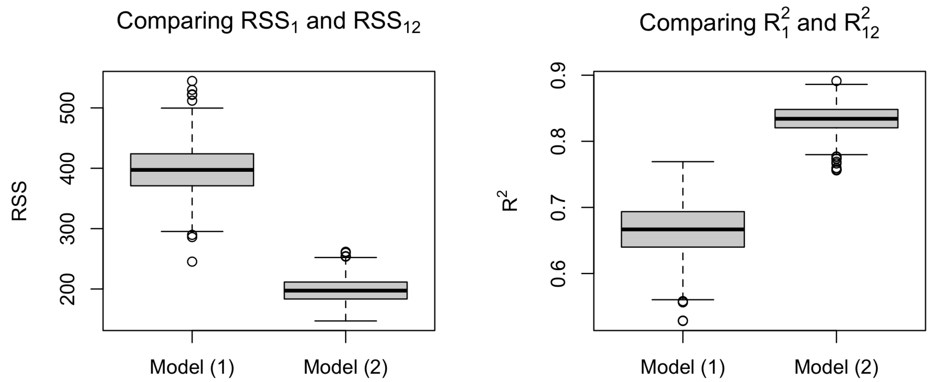

Figure 4: Comparing RSS and R^2 for model 1 and model 2

$$\stackrel{a.}{\leq} \min_{\beta_0,\beta_1,\beta_2=0} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 \cdot X_{i1} - 0 \cdot X_{i2})^2$$

$$= \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 \cdot X_{i1})^2$$

$$\stackrel{b.}{=} RSS_1$$

where $a.$ follows from that fact that if we pick an arbitrary β_2, we will always achieve a larger value than if we minimize over β_2 instead; $b.$ follows from rewriting RSS_1 as a minimization problem.

4. This question involves the use of multiple linear regression on the `Auto` data set, which is available as part of the `ISLR` library. **5+5+5 = 15 points**

   (a) Use the `lm()` function to perform a multiple linear regression with `mpg` as the response and all other variables except `name` as the predictors. Use the `summary()` function to print the results. Comment on the output. For instance:

      i. Is there a relationship between the predictors and the response?

      ii. Which predictors appear to have a statistically significant relationship to the response?

      iii. Provide an interpretation for the coefficient associated with the variable `year`.

   Make sure that you treat the qualitative variable `origin` appropriately.

   We fit the multiple linear model with all covariates (except name) as predictors for `mpg`. For the variable `origin`, we use *American* (`origin` = 1) as the baseline and produced two dummy

	$\hat{\beta}$	$\widehat{SE}(\hat{\beta})$	t statistic	p-value
(Intercept)	-17.955	4.677	-3.839	0.000
cylinders	-0.490	0.321	-1.524	0.128
displacement	0.024	0.008	3.133	0.002
horsepower	-0.018	0.014	-1.326	0.185
weight	-0.007	0.001	-10.243	0.000
acceleration	0.079	0.098	0.805	0.421
year	0.777	0.052	15.005	0.000
originEuropean	2.630	0.566	4.643	0.000
originJapanese	2.853	0.553	5.162	0.000

variables: `originEuropean` (1 if `origin` = 2, 0 otherwise) and `originJapanese` (1 if `origin` = 3, 0 otherwise).

The multiple linear regression model indicates that there is a negative association between `mpg` and `cylinders`, `horsepower`, and `weight`, whereas the relationship is positive between `mpg` and `displacement`, `acceleration`, `year`, `originEuropean`, `originJapanese`.

The following predictors have a statistically significant relationship (5% α level) to the response: `displacement`, `weight`, `year`, `originEuropean`, `originJapanese`.

The coefficient for the `year` variable suggests that with all other variables fixed, one unit increase in `year` (i.e., a newer model by one year) is associated with a 0.777 unit increase in `mpg`.

```
library(ISLR)
library(xtable) #if you are using latex
##
data(Auto)
Auto$origin <-  c("American", "European", "Japanese")[Auto$origin]
lm1 <- lm(mpg ~ . - name, data = Auto)
coefs <- data.frame(summary(lm1)$coefficients)
xtable(coefs,digits = 3) #generate Latex table
```

(b) Try out some models to predict `mpg` using **functions of** the variable `horsepower`. Comment on the best model you obtain. Make a plot with `horsepower` on the x-axis and `mpg` on the y-axis that displays both the observations and the fitted function (i.e. $\hat{f}$(`horsepower`)).

We start with a simple linear model and plot a) the fitted values and b) the fitted value versus residuals below. We note that while linear model is a decent fit, we tend to over-estimate observations with large mpg and under-estimate observations with small mpg.

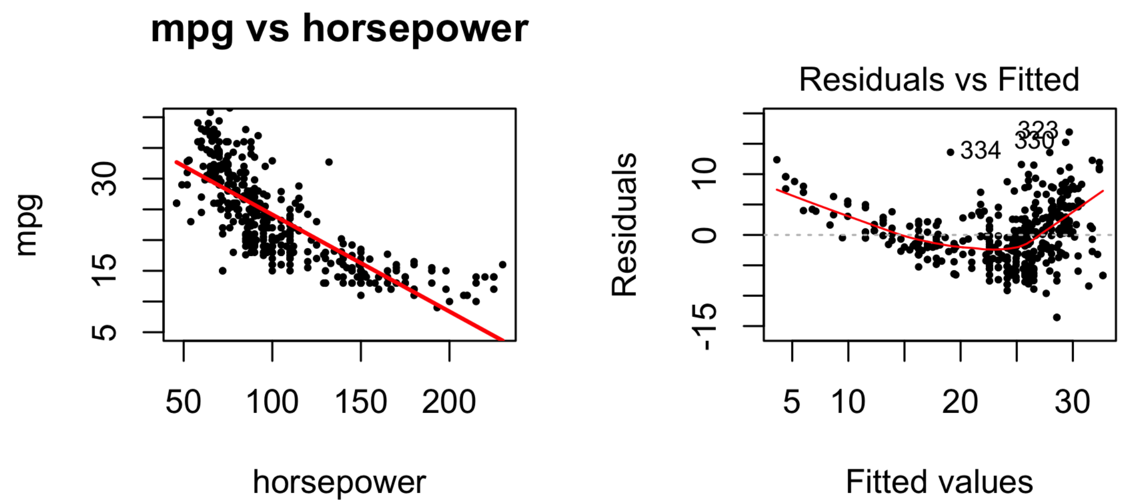

Figure 5: Fitted values and residual plot for the model $\texttt{mpg} = \beta_0 + \beta_1 \cdot \texttt{horsepower}$

We proceed with a more flexible quadratic model (indeed the scatter plot on the left indicates a quadratic model might be a better fit). We include the fitted values and residual plot below.

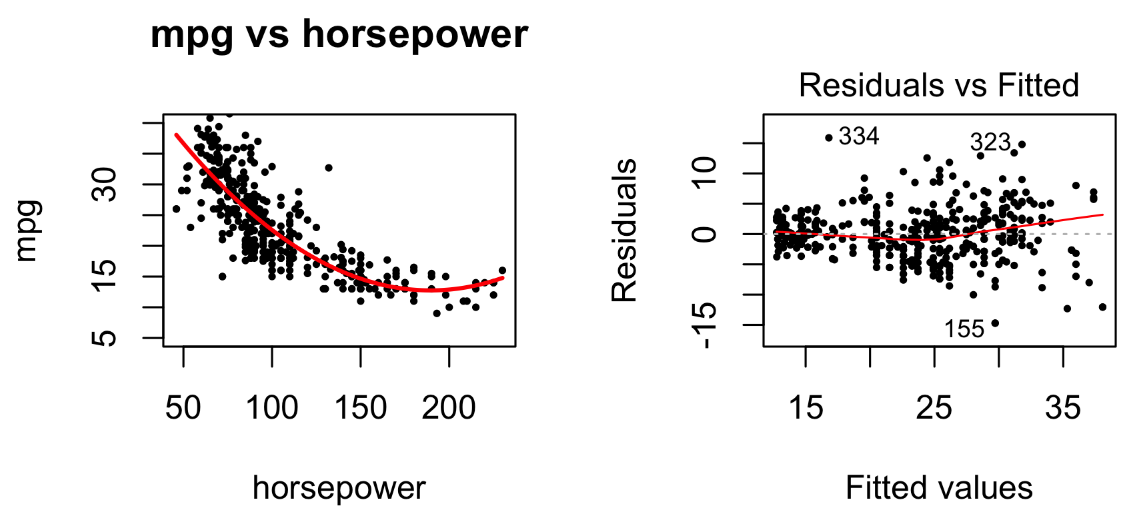

Figure 6: Fitted values and residual plot for the model $\texttt{mpg} = \beta_0 + \beta_1 \cdot \texttt{horsepower} + \beta_2 \cdot \texttt{horsepower}^2$

Our quadratic model improves the fit by quite a bit visually and the residuals are on average close to 0. Now one can proceed with more complex function forms (e.g. cubic function, $\log$ **function,**

etc.), the added terms most likely won't contribute much to the fit. For instance, the cubic term is not statistically significant and the fitted values are very similar to those in the quadratic model.

```
par(mfrow = c(1, 2))
fit0 <- lm(mpg ~ horsepower, data = Auto)
summary(fit0)
new <- data.frame(horsepower=c(min(Auto$horsepower):max(Auto$horsepower)))
plot(Auto$horsepower, Auto$mpg, xlab = "horsepower", ylab = "mpg",
     main = "mpg vs horsepower",ylim = c(5,40),pch=16,cex=0.5)
lines(new$horsepower, predict(fit0, new), col = "red", lwd = 2)
plot(fit0, which = 1,pch=16,cex=0.5)

par(mfrow = c(1, 2))
fit1 <- lm(mpg ~ horsepower+I(horsepower^2), data = Auto)
summary(fit1)
new <- data.frame(horsepower=c(min(Auto$horsepower):max(Auto$horsepower)))
plot(Auto$horsepower, Auto$mpg, xlab = "horsepower", ylab = "mpg",
     main = "mpg vs horsepower",ylim = c(5,40),pch=16,cex=0.5)
lines(new$horsepower, predict(fit1, new), col = "red", lwd = 2)
plot(fit1, which = 1,pch=16,cex=0.5)

fit2 <- lm(mpg ~ horsepower+I(horsepower^2)+I(horsepower^3), data = Auto)
summary(fit2)
```

(c) Now fit a model to predict `mpg` using `horsepower`, `origin`, and an interaction between `horsepower` and `origin`. Make sure to treat the qualitative variable `origin` appropriately. Comment on your results. Provide a careful interpretation of each regression coefficient.

	Estimate	Std..Error	t.value	Pr...t..
(Intercept)	34.476	0.891	38.709	0.000
horsepower	-0.121	0.007	-17.099	0.000
originEuropean	10.997	2.396	4.589	0.000
originJapanese	14.340	2.464	5.819	0.000
horsepower:originEuropean	-0.101	0.028	-3.626	0.000
horsepower:originJapanese	-0.109	0.029	-3.752	0.000

To interpret each of the coefficients,

- **On average, an American vehicle with 0 horsepower is expected to have `mpg` to be 34.476 (0 horsepower, of course, is unrealistic and therefore intercept in a regression model is often not very interpretable by itself).)**
- **For an American vehicle, an unit increase in engine horsepower is associated with a 0.121 decrease in `mpg`.**

- **For a European vehicle, an unit increase in engine horsepower is associated with a $0.121 + 0.101 = 0.222$ decrease in `mpg`.**
- **For a Japanese vehicle, an unit increase in engine horsepower is associated with a $0.121 + 0.109 = 0.230$ decrease in `mpg`.**
- **For vehicles with the same horsepower, European vehicles are expected to have 10.997 higher mpg than American vehicles on average, and Japanese vehicles are expected to have 14.340 higher mpg than American vehicles on average.**

```
fit.c <- lm(mpg ~ horsepower * origin, data = Auto)
coefs <- data.frame(summary(fit.c)$coefficients)
xtable(coefs,digits = 3) #generate Latex table
```

5. Consider fitting a model to predict credit card `balance` using `income` and `student`, where `student` is a qualitative variable that takes on one of three values: `student`$\in$ {`graduate`, `undergraduate`, `not student`}. **5+5+5+5+5 = 25 points**

   (a) Encode the student variable using two dummy variables, one of which equals 1 if `student=graduate` (and 0 otherwise), and one of which equals 1 if `student=undergraduate` (and 0 otherwise). Write out an expression for a linear model to predict `balance` using `income` and `student`, using this coding of the dummy variables. Interpret the coefficients in this linear model.

   This model would be:

   $$\texttt{balance} = \beta_0 + \beta_1 \cdot \texttt{income} + \beta_2 \cdot 1\{\texttt{student=graduate}\} + \beta_3 \cdot 1\{\texttt{student=undergraduate}\}$$

   To interpret each of the coefficients,

   - **β_0 represents the average credit card balance for non-students with 0 income.**
   - **β_1 represents the difference in average credit card balance associated with a one-unit increase in income, comparing subjects with the same student status.**
   - **β_2 represents the difference in average credit card balance comparing graduate students with non-students who have equal incomes.**
   - **β_3 represents the difference in average credit card balance comparing undergraduate students with non-students who have equal incomes.**

   (b) Now encode the student variable using two dummy variables, one of which equals 1 if `student=not student` (and 0 otherwise), and one of which equals 1 if `student=graduate` (and 0 otherwise). Write out an expression for a linear model to predict `balance` using `income` and `student`, using this coding of the dummy variables. Interpret the coefficients in this linear model.

This model would be:

$$\texttt{balance} = \beta_0 + \beta_1 \cdot \texttt{income} + \beta_2 \cdot 1\{\texttt{student=graduate}\} + \beta_3 \cdot 1\{\texttt{student=not student}\}$$

We interpret the coefficients here very similarly. The only difference is in our reference group for education:

- **β_0 represents the average credit card balance for undergraduates with 0 income.**
- **β_1 represents the difference in average credit card balance associated with a one-unit increase in income, comparing subjects with the same student status.**
- **β_2 represents the difference in average credit card balance comparing graduate students with undergraduates who have equal incomes.**
- **β_3 represents the difference in average credit card balance comparing undergraduate students with undergraduates who have equal incomes.**

(c) Using the coding in (a), write out an expression for a linear model to predict `balance` using `income`, `student`, and an interaction between `income` and `student`. Interpret the coefficients in this model.

This model would be:

$$\texttt{balance} = \beta_0 + \beta_1 \cdot 1\{\texttt{income}\} + \beta_2 \cdot \{\texttt{student=graduate}\} + \beta_3 \cdot \{\texttt{student=undergraduate}\}$$

$$+ \beta_4 \cdot \texttt{income} \cdot 1\{\texttt{student=graduate}\} + \beta_5 \cdot \texttt{income} \cdot 1\{\texttt{student=undergraduate}\}$$

To interpret each of the coefficients,

- **β_0 represents the average credit card balance for non-students with 0 income.**
- **β_1 represents the difference in average credit card balance associated with a one-unit increase in income, comparing non-students only.**
- **Now, we can write the difference in average credit card balance comparing comparing graduate students with non-students who have equal incomes as $\beta_2 + \beta_4 \cdot \texttt{income}$.**
- **Similarly, we can write the difference in average credit card balance comparing comparing undergraduate students with non-students who have equal incomes as $\beta_3 + \beta_5 \cdot \texttt{income}$.**

- **We can also write the difference in average credit card balance associated with a one-unit increase in income, comparing graduate students only, as $\beta_1 + \beta_4$. Comparing this with the interpretation of β_1 above, we can consider β_4 as the additional difference in credit card balance by an unit increase in income by being a graduate student vs. a non-student.**
- **Similarly, we can write the difference in average credit card balance associated with a one-unit increase in income, comparing undergraduate students only, as $\beta_1 + \beta_5$. Comparing this with the interpretation of β_1 above, we can consider β_5 as the additional difference in credit card balance by an unit increase in income by being an undergraduate student vs. a non-student.**

(d) Using the coding in (b), write out an expression for a linear model to predict `balance` using `income`, `student`, and an interaction between `income` and `student`. Interpret the coefficients in this model.

$$\texttt{balance} = \beta_0+\beta_1\cdot\texttt{income}+\beta_2\cdot 1\{\texttt{student=graduate}\}+\beta_3\cdot 1\{\texttt{student=non student}\}$$

$$+\beta_4\cdot\texttt{income}\cdot 1\{\texttt{student=graduate}\}+\beta_5\cdot\texttt{income}\cdot 1\{\texttt{student=non student}\}$$

We interpret the coefficients here very similarly. The only difference is in our baseline (or reference) group:

- **β_0 represents the average credit card balance for undergraduates with 0 income.**
- **β_1 represents the difference in average credit card balance associated with a one-unit increase in income, comparing undergraduates only.**
- **Now, we can write the difference in average credit card balance comparing comparing graduate students with undergraduates who have equal incomes as $\beta_2 + \beta_4$`income`. Therefore, β_2 and β_4 represent the intercept and slope of the line which defines this difference.**
- **Similarly, we can write the difference in average credit card balance comparing comparing non-students students with undergraduates who have equal incomes as $\beta_3 + \beta_5$`income`. Therefore, β_3 and β_5 represent the intercept and slope of the line which defines this difference.**
- **We can also write the difference in average credit card balance associated with a one-unit increase in income, comparing graduate students only, as $\beta_1 + \beta_4$. Comparing this with the interpretation of β_1 above, we can consider β_4 as the additional difference in credit card balance by an unit increase in income by being a graduate student vs an undergraduate.**

- **Similarly, we can write the difference in average credit card balance associated with a one-unit increase in income, comparing non-students only, as $\beta_1+\beta_5$. Comparing this with the interpretation of β_1 above, we can consider β_5 as the additional difference in credit card balance by an unit increase in income by being a non-student vs an undergraduate student.**

(e) Using simulated data for `balance`, `income`, and `student`, show that the fitted values (predictions) from the models in (a)–(d) do not depend on the coding of the dummy variables (i.e. the models in (a) and (b) yield the same fitted values, as do the models in (c) and (d)).

We used the code below to generate data and compare fitted models:

```
set.seed(1234)
income <- rchisq(1000, df = 1000)
student <- sample(c("graduate", "undergraduate", "not student"),
    size = 1000, replace = TRUE)
balance <- 300 + 0.1*income + 3*(student=="graduate") +
    4*(student=="undergraduate") + rnorm(100)

m1 <- lm(balance ~ income + (student=="graduate") +
    (student=="undergraduate"))
m2 <- lm(balance ~ income + (student=="graduate") +
    (student=="not student"))
plot(x=fitted(m1), y=fitted(m2),
     pch=16,
     cex=0.6,
     xlab = 'Fitted values from model (a)',
     ylab = 'Fitted values from model (b)',
     main = 'Comparing fitted values from model (a) and (b) (y=x plotted)')
abline(0,1)

m3 <- lm(balance ~ income + (student=="graduate") +
    (student=="undergraduate") +
    income*(student=="graduate") +
    income*(student=="undergraduate"))
m4 <- lm(balance ~ income + (student=="graduate") +
    (student=="undergraduate") +
    income*(student=="graduate") +
    income*(student=="not student"))
plot(x=fitted(m3), y=fitted(m4),
     pch=16,
     cex=0.6,
     xlab = 'Fitted values from model (c)',
```

```
ylab = 'Fitted values from model (d)',
main = 'Comparing fitted values from model (c) and (d) (y=x plotted)')
```

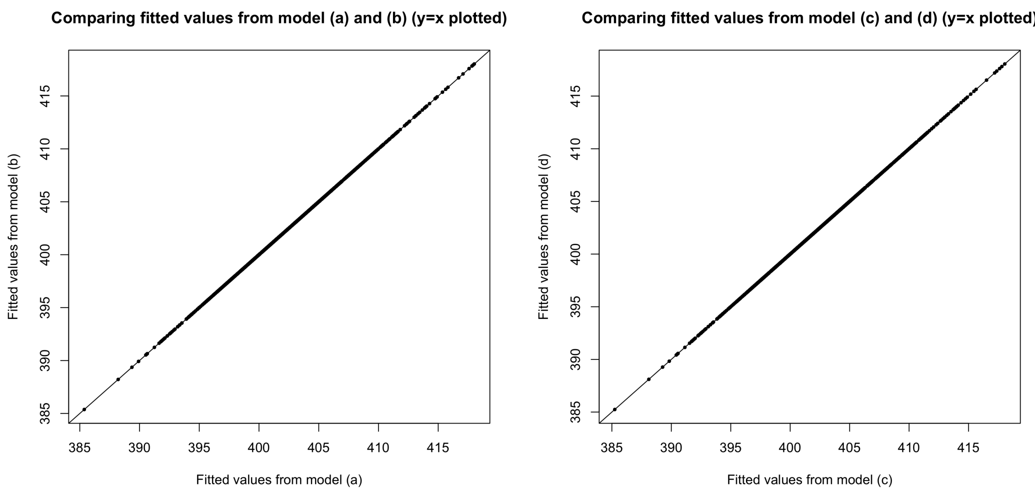

Figure 7: Comparing fitted model values

6. This problem has to do with logistic regression. **5+5 = 10 points**

   (a) Suppose you fit a logistic regression to some data and find that for a given observation $x = (x_1, \ldots, x_p)^T$, the estimated log-odds equals 0.23. What is $P(Y = 1 \mid X = x)$?

   Recall that in logistic regression, we model the log-odds as a linear combination of predictors, i.e.,

$$\log(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)} = \sum_{i=1}^{p} x_i \cdot \hat{\beta}_p = 0.23$$

   . This implies that

$$P(Y=1|X=x) = \frac{\exp(\sum_{i=1}^{p} x_i \cdot \hat{\beta}_p)}{1+\exp(\sum_{i=1}^{p} x_i \cdot \hat{\beta}_p)}$$

   . Plugging in 0.23 gives us

$$P(Y=1|X=x) = \frac{e^{0.23}}{1+e^{0.23}} \approx 0.557$$

   (b) In the same setting as (a), suppose you are now interested in the observation $x^* = (x_1 + 0.5, x_2 - 5, x_3, x_4, \ldots, x_p)^T$. In other words, $x_1^* = x_1 + 0.5$, $x_2^* = x_2 - 5$, and $x_j^* = x_j$ for $j \geq 3$. Write out a simple expression for

$P(Y = 1 \mid X = x^*)$. Your answer will involve the estimated coefficients in the logistic regression model, as well as the number 0.23.

Since

$$P(Y=1|X=x^*) = \frac{\exp(\sum_{i=1}^{p} x_i^* \cdot \hat{\beta}_p)}{1+\exp(\sum_{i=1}^{p} x_i^* \cdot \hat{\beta}_p)}$$

, it suffices to understand how $\sum_{i=1}^{p} x_i^* \cdot \hat{\beta}_p$ is related to $\sum_{i=1}^{p} x_i \cdot \hat{\beta}_p$.

$$\sum_{i=1}^{p} x_i^* \cdot \hat{\beta}_p = (\sum_{i=1}^{p} x_i \cdot \hat{\beta}_p) + 0.5 \cdot \hat{\beta}_1 - 5 \cdot \hat{\beta}_2$$

.

Plugging it into the formula above gives us

$$P(Y=1|X=x^*) = \frac{\exp\left((\sum_{i=1}^{p} x_i \cdot \hat{\beta}_p) + 0.5 \cdot \hat{\beta}_1 - 5 \cdot \hat{\beta}_2\right)}{1+\exp\left((\sum_{i=1}^{p} x_i \cdot \hat{\beta}_p) + 0.5 \cdot \hat{\beta}_1 - 5 \cdot \hat{\beta}_2\right)}$$

$$= \frac{\exp(0.23) \cdot \exp(0.5 \cdot \hat{\beta}_1 - 5 \cdot \hat{\beta}_2)}{1+\exp(0.23) \cdot \exp(0.5 \cdot \hat{\beta}_1 - 5 \cdot \hat{\beta}_2)}$$