Homework 04 Solutions

9/27/2021

1.

(1a)

We will use several facts from lecture. First, let $W \sim N(\mu, \Sigma)$ be an arbitrary multivariate normal random vector and A a matrix whose second dimension is the same as the length of W, so that AW is defined. Then we have

$$AW \sim N(A\mu, A\Sigma A^T) \tag{1}$$

Additionally, we have from lecture that

$$HX\beta = X\beta \tag{2}$$
$$(I-H)(I-H)^T = (I-H)(I-H) = (I-H) \tag{3}$$
$$y \sim N(X\beta, \sigma^2 I) \tag{4}$$

Let $\hat{y}$ be the $n \times 1$ column vector of fitted values.

$$\begin{aligned}
\hat{e} &= y - \hat{y} \\
&= y - Hy \\
&= (I-H)y \\
&\sim N((I-H)X\beta, (I-H)\sigma^2 I(I-H)^T) && \text{(by facts (1) and (4))} \\
&= N(0, \sigma^2(I-H)(I-H)^T) && \text{(by fact (2))} \\
&= N(0, \sigma^2(I-H)) && \text{(by fact (3))}
\end{aligned}$$

□

(1b)

(i)

In the intercept-only model, the design matrix X is an $n \times 1$ matrix in which each entry is 1. This can be used to derive the hat matrix H_1:

$$\begin{aligned}
X^T X &= [n]_{1\times 1} && \text{(the } 1\times 1 \text{ matrix whose only entry is } n\text{)} \\
\implies (X^T X)^{-1} &= [1/n]_{1\times 1} && \text{(the } 1\times 1 \text{ matrix whose only entry is } 1/n\text{)} \\
\implies X(X^T X)^{-1} X^T &= \frac{1}{n} X X^T \\
&= \frac{1}{n} \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix}_{n\times n} \\
&= H_1 && \text{(by the definition of the hat matrix)}
\end{aligned}$$

For any i in $1, 2, \dots, n$, we derive $\hat{y}_i$ by multiplying the ith row of H_1 by the vector y. In other words, for the intercept-only model, $\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$ for all i. □

(ii)

To avoid using $\bar{y}$ to denote both a scalar and a vector, let's immediately use the fact that $H_1 y$ is equal to the $n \times 1$ vector $\begin{bmatrix}\bar{y} & \bar{y} & \dots & \bar{y}\end{bmatrix}^T$, as shown in the previous problem. Rewriting the covariance in vector form, per the problem suggestion, we have:

$$\begin{aligned}
\text{Cov}(y, \hat{y}) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y}) \\
&= \frac{1}{n}(y - H_1 y)^T(\hat{y} - H_1 y) \\
&= \frac{1}{n}(y - H_1 y)^T(Hy - H_1 y) \\
&= \frac{1}{n}[(I - H_1)y]^T(H - H_1)y \\
&= \frac{1}{n}y^T(I - H_1)^T(H - H_1)y && \text{(since } (AB)^T = B^T A^T\text{)} \\
&= \frac{1}{n}y^T(H - H_1 - H_1^T H + H_1^T H_1)y \\
&= \frac{1}{n}y^T(H - H_1 H)y && \text{(since } H_1^T = H_1 \text{ and } H_1 H_1 = H_1\text{)} \\
&= \frac{1}{n}y^T(H - H_1)y && \text{(since } H_1 H = H H_1 = H_1\text{)}
\end{aligned}$$

□

(iii)

Recall that the sample correlation is the standardized sample covariance, defined as

$$\begin{aligned}
widehat\text{Corr}(y, \hat{y}) &= \frac{\hat{\text{Cov}}(y, \hat{y})}{\sqrt{\hat{\sigma}_y^2 \hat{\sigma}_{\hat{y}}^2}} \\
&= \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}} \\
&= \frac{(y - H_1 y)^T(\hat{y} - H_1 y)}{\sqrt{(y - H_1 y)^T(y - H_1 y)(\hat{y} - H_1 y)^T(\hat{y} - H_1 y)}} \\
&= \frac{y^T(H - H_1)y}{\sqrt{y^T(I - H_1)^T(I - H_1)y y^T(H - H_1)^T(H - H_1)y}} && \text{(using the previous problem)} \\
&= \frac{\sqrt{y^T(H - H_1)y}}{\sqrt{y^T(I - H_1)^T(I - H_1)y}} \\
&= \frac{\sqrt{y^T(H - H_1)y}}{\sqrt{y^T(I - H_1)y}} && \text{(using a version of fact (3))} \\
&= \sqrt{\frac{SS_{reg}}{SST}} \\
&= \sqrt{R^2} && \text{(by the definition of } R^2\text{)}
\end{aligned}$$

And therefore, $R^2 = \widehat{\text{Corr}}(y, \hat{y})^2$. □

(1c)

The sample correlation is 0 iff the sample covariance is 0, so let's just concern ourselves with the sample covariance and not worry about the denominator in the correlation.

Recall that the residuals sum to 0 by construction, which also means that their sample mean is 0. Hence:

$$\begin{aligned}
\hat{\text{Cov}}(\hat{e}, \hat{y}) &= \frac{1}{n}\sum_{i=1}^{n}(\hat{e}_i - 0)(\hat{y}_i - \bar{y}) \\
&= \frac{1}{n}(\hat{e})^T(\hat{y} - \bar{y}) \\
&= \frac{1}{n}(y - \hat{y})^T(\hat{y} - \bar{y}) \\
&= \frac{1}{n}[(I - H)y]^T(H - H_1)y \\
&= \frac{1}{n}y^T(I - H)(H - H_1)y && \text{(since } (I-H)^T = (I-H)\text{)} \\
&= \frac{1}{n}y^T(H - H_1 - HH + HH_1)y \\
&= \frac{1}{n}y^T(H - H_1 - H + H_1)y && \text{(since } HH_1 = H_1\text{)} \\
&= 0
\end{aligned}$$

□

2. (Sheather, Ch 5, pp. 146–147, #1, with extra questions.)

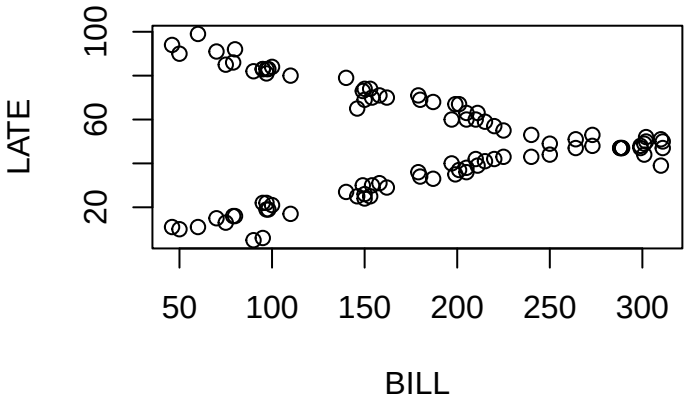
(2a) Develop model to predict LATE from BILL.

Let's start with a little EDA...

```
overdue <- read.table("overdue.txt",header=T)
str(overdue)
```

```
## 'data.frame':    96 obs. of  2 variables:
##  $ LATE: int  16 47 22 47 47 21 44 27 19 48 ...
##  $ BILL: int  79 264 97 289 288 100 250 140 97 299 ...
```

```
plot(LATE ~ BILL, data=overdue)
```



From the scatter plot, it looks like regression is going to fail, because LATE is not a single-valued function of BILL. Indeed, if we do a simple regression of LATE on BILL, we get

```
round(summary(lm(LATE ~ BILL,data=overdue))$coefficients,2)
```

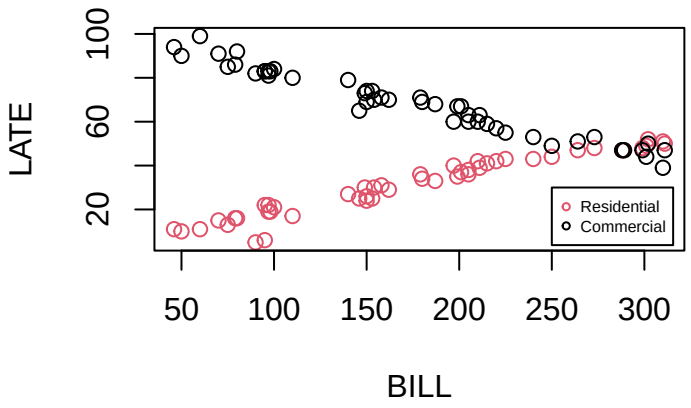
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    51.98       5.96    8.72     0.00
## BILL           -0.01       0.03   -0.40     0.69
```

and the coefficient on BILL is clearly not significantly different from zero.

Looking back at the problem statement, we also know that the first 48 observations are for residential accounts, and the last 48 are for commercial accounts. Maybe this helps explain the structure we see above?

So, let's create a dummy variable for "residential", and colored the points according to this dummy variable.

```
overdue$residential <- c(rep(1,48),rep(0,48))
plot(LATE ~ BILL, data=overdue,col=residential + 1)
legend(250,30,legend=c("Residential","Commercial"),pch=1,col=c(2,1),cex=0.5)
```

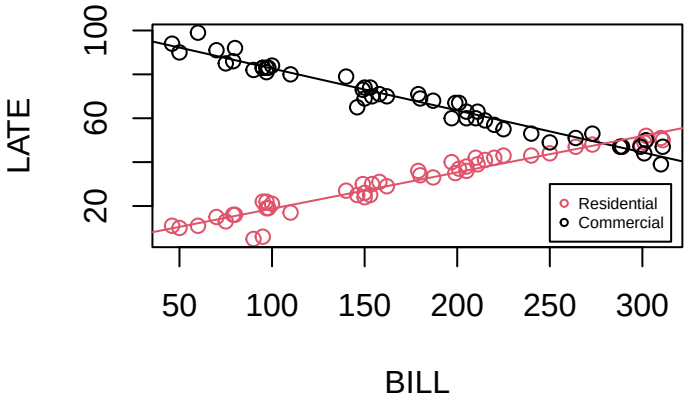


Much better: Now we clearly see that the two "branches" in the scatter plot correspond to the two types of account, with different slopes and intercepts for each type of account. This suggests we should fit an ANCOVA model with interactions:

```
lm.2a <- lm(LATE ~ BILL*residential, data=overdue)
print(coefs <- coef(lm.2a))
```

```
##      (Intercept)             BILL      residential BILL:residential
##      101.7581844       -0.1909615      -99.5485607        0.3566445
```

```
plot(LATE ~ BILL, data=overdue,col=residential + 1)
abline(a=coefs[1], b=coefs[2],col=1)
abline(a=coefs[1]+coefs[3], b=coefs[2]+coefs[4],col=2)
legend(250,30,legend=c("Residential","Commercial"),pch=1,col=c(2,1),cex=0.5)
```



The new scatter plot suggests that this may be pretty good model for the data.

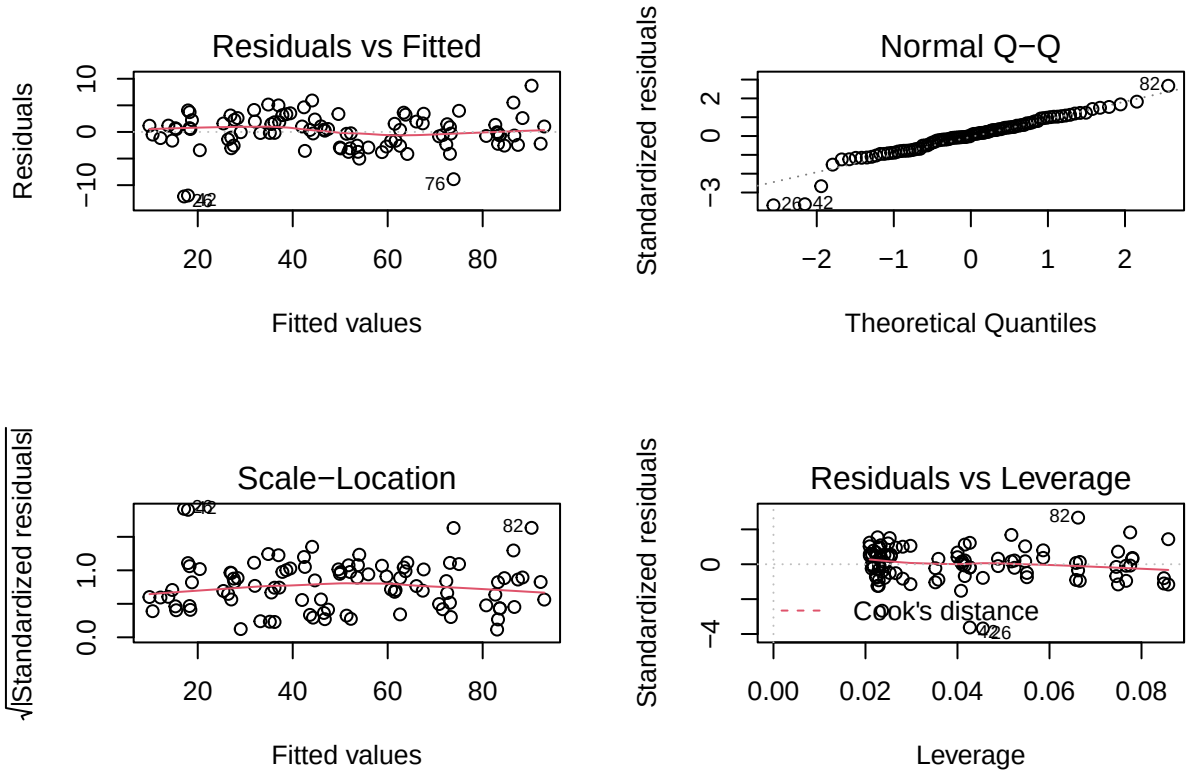
(2b) Summary & Residual Diagnostics.

Here is the model summary and the set of four standard diagnostic plots:

```
summary(lm.2a)
```

```
##
## Call:
## lm(formula = LATE ~ BILL * residential, data = overdue)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.1211  -2.2163   0.0974   1.9556   8.6995
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      101.758184   1.198504   84.90   <2e-16 ***
## BILL              -0.190961   0.006285  -30.38   <2e-16 ***
## residential      -99.548561   1.694940  -58.73   <2e-16 ***
## BILL:residential   0.356644   0.008888   40.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.371 on 92 degrees of freedom
## Multiple R-squared:  0.9803, Adjusted R-squared:  0.9796
## F-statistic:  1524 on 3 and 92 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(lm.2a)
```

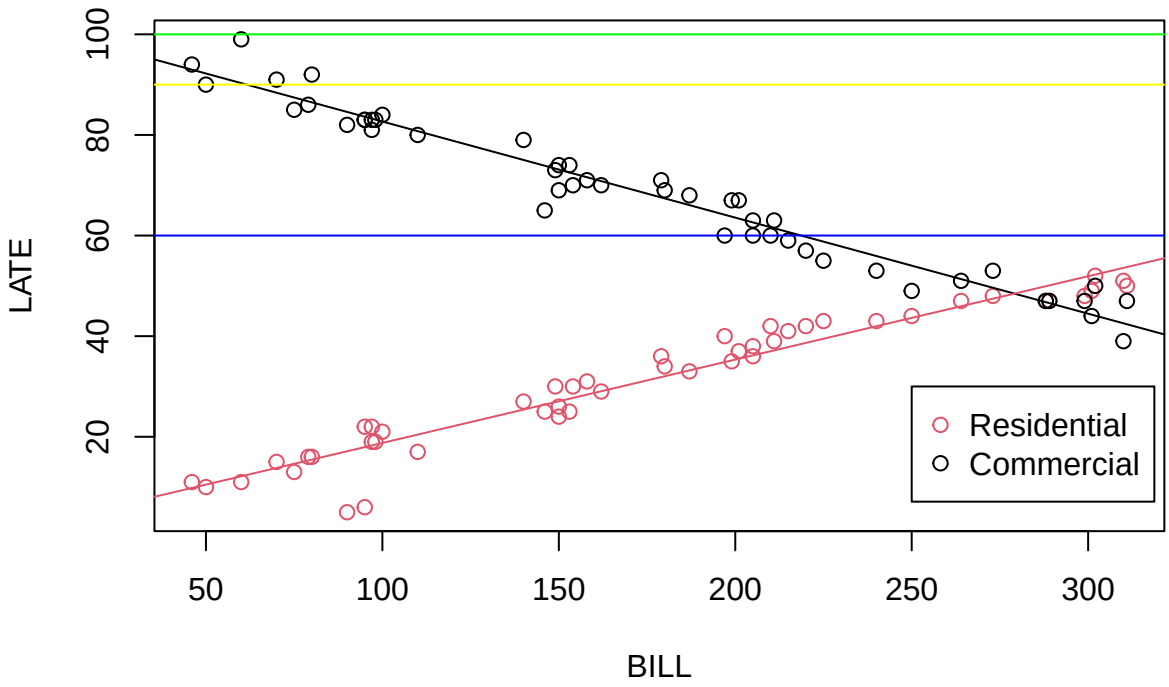


This seems like a pretty good model. The regression coefficients are all strongly significantly different from zero, and the $R^2 = 0.98$. Looking at the residual diagnostic plots, we see that the residuals are centered at zero and show no vertical patterns; the QQ plot suggests that the residuals are approximately normal with just 3 or 4 outliers, the Scale-Location plot is consistent with constant variance, and there aer no points with Cook's distance above 0.5. The points with high leverage are generally not outliers, and the points that are outliers have modest leverage.

(2c) "Under 60 days or your money back?"

To think about the "Under 60 days or your money back!" slogan, let's look at the final scatter plot with the fitted regression model again, but with horizontal lines at 60, 90 and 100 days overlaid:

```
plot(LATE ~ BILL, data=overdue,col=residential + 1)
abline(a=coefs[1], b=coefs[2],col=1)
abline(a=coefs[1]+coefs[3], b=coefs[2]+coefs[4],col=2)
abline(h=60,col="Blue")
abline(h=90,col="Yellow")
abline(h=100,col="Green")
legend(250,30,legend=c("Residential","Commercial"),pch=1,col=c(2,1))
```



It looks like 60 days is enough for residential accounts (even if we extrapolate a little to the right of the data using the regression line), but we need up to 100 days for commercial accounts.

If we want the slogan to work for almost all collections, perhaps we should recommend

- "Under 90 days or your money back!" or perhaps
- "Under 100 days or your money back!"

since 90 days is enough for most collections, even commercial ones, and 100 days is enough for viturally all of them.

On the other hand, if we can restrict the advertising to Residential accounts only, then "Under 60 days" seems just fine: Based on this plot, we would expect virtually all Residential account to be collected in 60 days or less.

3. (Sheather, Ch 5, p. 147, #2.)

We begin with a little exploration, and finding a good model.

```
houston <- read.csv("HoustonChronicle.csv",header=T)
str(houston)
```

```
## 'data.frame':    122 obs. of  5 variables:
##  $ District             : chr  "Alvin" "Alvin" "Angleton" "Angleton" ...
##  $ X.Repeating.1st.Grade: num  4.1 5.8 7.1 6.7 7.3 2.6 8.2 2.3 12.5 0 ...
##  $ X.Low.income.students: num  49.7 41.1 44.2 30.2 49.4 33.7 45.6 29.7 71.7 37.6 ...
##  $ Year                 : int  2004 1994 2004 1994 2004 1994 2004 1994 2004 1994 ...
##  $ County               : chr  "Brazoria" "Brazoria" "Brazoria" "Brazoria" ...
```

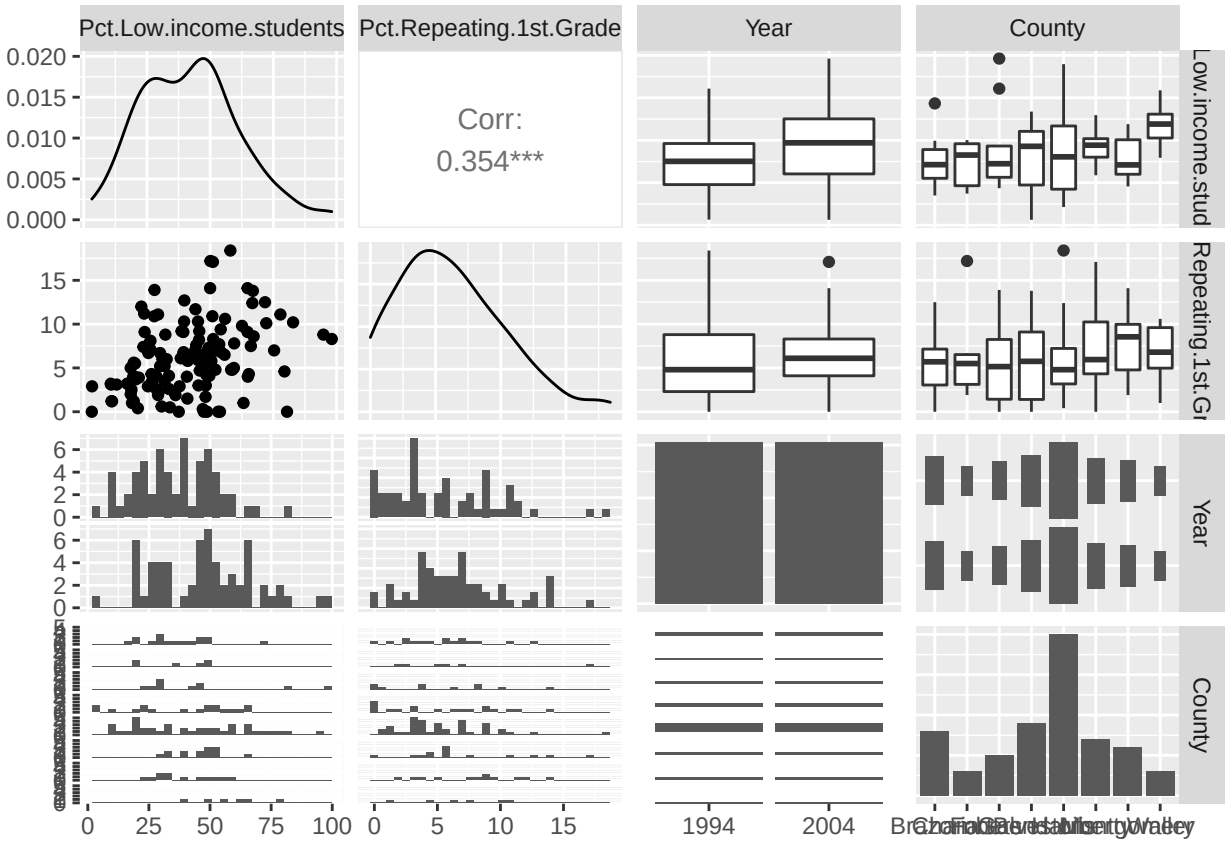
```
cat("\n")
```

```
for (i in names(houston)) {
  cat(i,"\n")
  print(summary(houston[,i]))
  cat("Number of unique values:",length(unique(houston[,i])),"\n\n")
}
```

```
## District
##    Length     Class      Mode
##       122 character character
## Number of unique values: 61
##
## X.Repeating.1st.Grade
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.100   5.700   6.076   8.750  18.400
## Number of unique values: 77
##
## X.Low.income.students
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.20   27.15   41.35   41.88   53.02   98.10
## Number of unique values: 111
##
## Year
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1994    1994    1999    1999    2004    2004
## Number of unique values: 2
##
## County
##    Length     Class      Mode
##       122 character character
## Number of unique values: 8
```

Apparently, there are only two "year" values: 1994 and 2004; so we will convert this to a factor variable, just to make interpretation of the coefficients easier; and we make a scatterplot matrix ("pairs" plot) of all of the variables except district. I'm also going to rename the two continuous variables to something more suggestive, and re-order the variables, so that the "pairs" plot makes a little more sense.

```
houston$Year <- as.factor(houston$Year)
names(houston)[c(2,3)] <- c("Pct.Repeating.1st.Grade", "Pct.Low.income.students")
houston <- houston[,c(1,3,2,4,5)]
ggpairs(houston[,-1])  # there are too many districts for the plot, so we omit it
```



There is some evidence in the "pairs" plots for all three of the hypotheses in parts (a), (b) and (c), but we now fit regression and ANCOVA models to see which effects are "significant" (i.e. large enough that they are not likely just due to noise in the data).

(3a) Low income associated with repeating first grade?

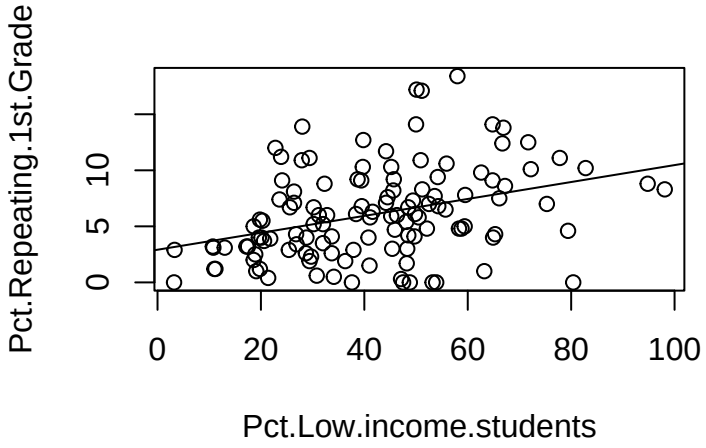
Let's start with a simple regression of repeating first grade on low income

```
summary(lm.3a <- lm(Pct.Repeating.1st.Grade ~ Pct.Low.income.students,data=houston))
```

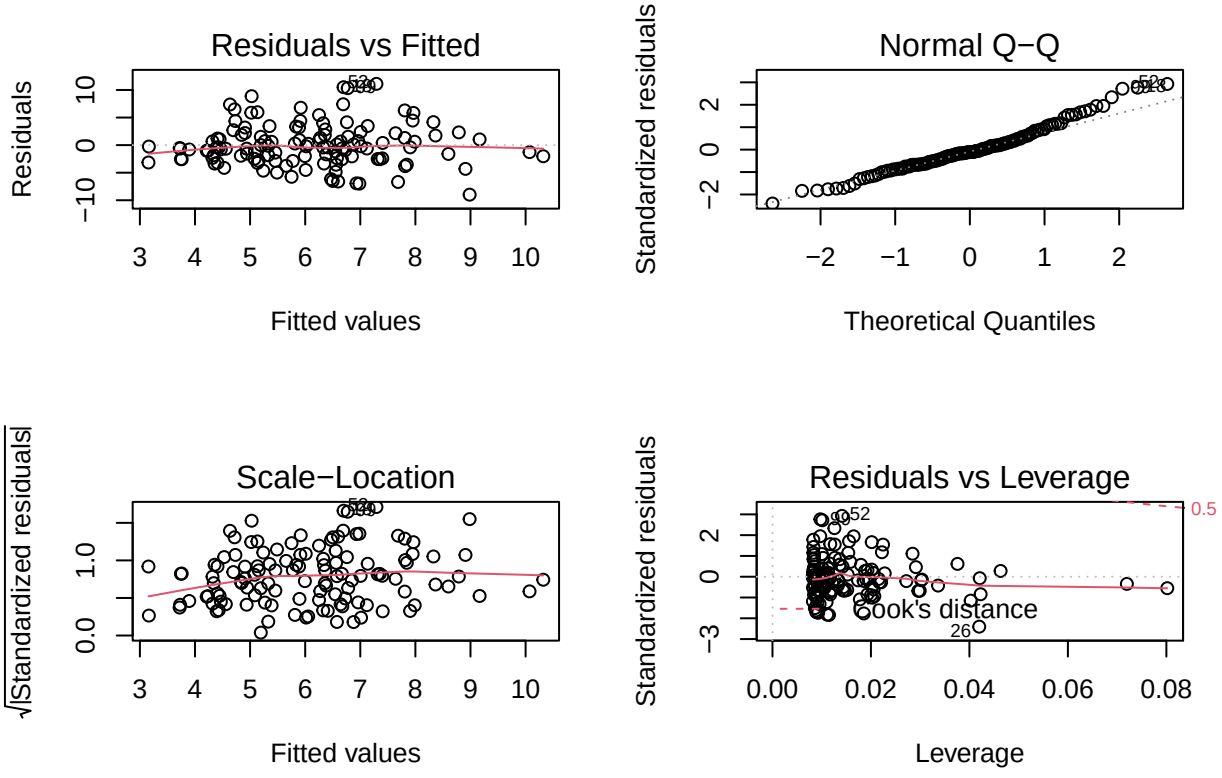
```
##
## Call:
## lm(formula = Pct.Repeating.1st.Grade ~ Pct.Low.income.students,
##     data = houston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9845 -2.5072 -0.4184  1.8505 11.1067
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.91419    0.83836   3.476 0.000709 ***
## Pct.Low.income.students   0.07550    0.01823   4.141 6.47e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.821 on 120 degrees of freedom
## Multiple R-squared:  0.125,  Adjusted R-squared:  0.1177
```

```
## F-statistic: 17.14 on 1 and 120 DF,  p-value: 6.472e-05
```

```
par(mfrow=c(1,1))
plot(Pct.Repeating.1st.Grade ~ Pct.Low.income.students,data=houston)
abline(lm.3a)
```




```
par(mfrow=c(2,2))
plot(lm.3a)
```



This actually looks like a good regression; the summary shows a very significant increasing association between percent of students repeating first grade and percent of students with low (family) incomes: the coefficient on percent low income students is $\hat{\beta}_1 = 0.07550$, with $SE(\hat{\beta}_1) = 0.01823$. The effect seems rather small, however—we expect an increase of only 0.08% of kids repeating first grade, for every 1% increase in kids in poverty.

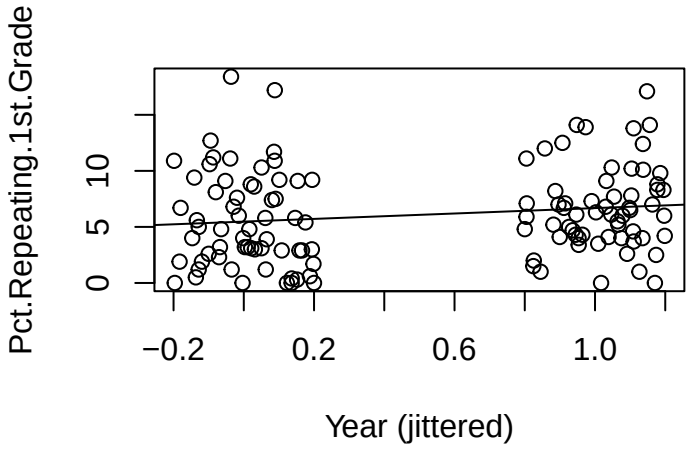
(3b) More students repeating first grade in 2004-2005 than in 1994-1995?

For this question, we will just regress repeating first grade on year

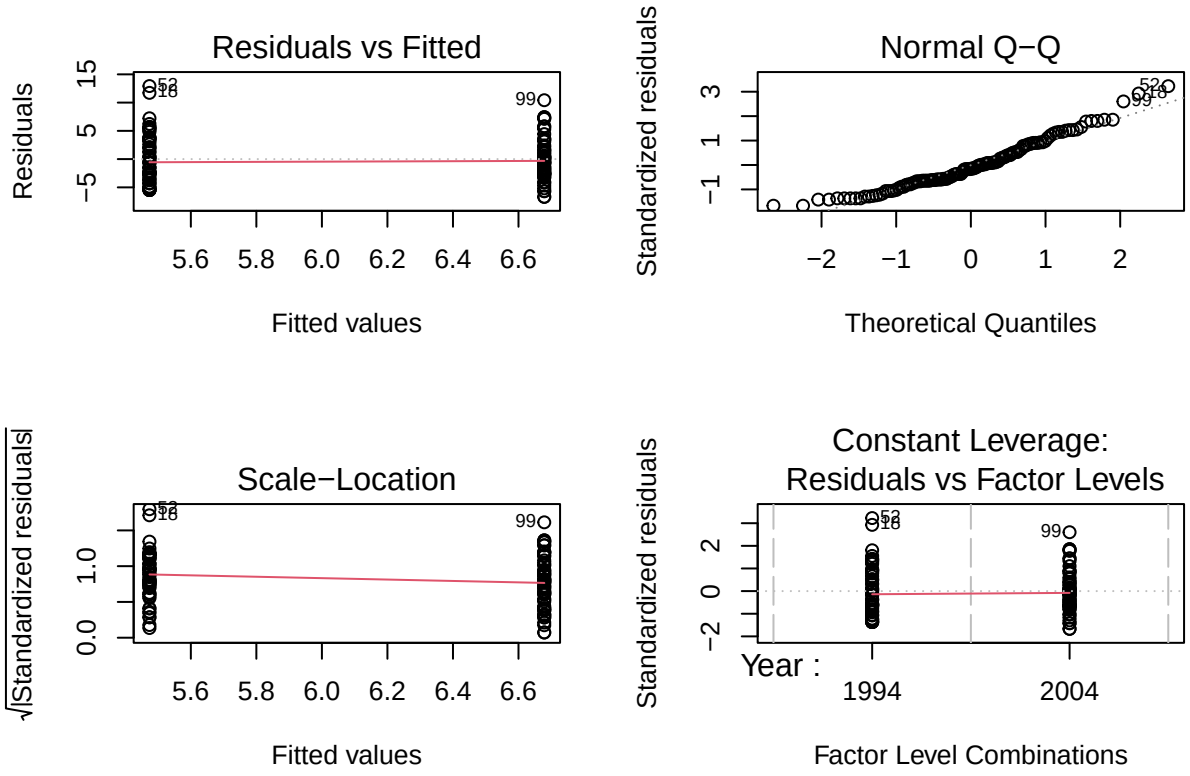
```
summary(lm.3b <- lm(Pct.Repeating.1st.Grade ~ Year,data=houston))
```

```
##
## Call:
## lm(formula = Pct.Repeating.1st.Grade ~ Year, data = houston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6787 -2.6537 -0.6262  2.5750 12.9262
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.4738     0.5172  10.584   <2e-16 ***
## Year2004      1.2049     0.7314   1.647    0.102
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.039 on 120 degrees of freedom
## Multiple R-squared:  0.02212,    Adjusted R-squared:  0.01397
## F-statistic: 2.714 on 1 and 120 DF,  p-value: 0.1021
```

```
par(mfrow=c(1,1))
jy <- jitter(as.numeric(houston$Year)-1)
plot(Pct.Repeating.1st.Grade ~ jy,data=houston,xlab="Year (jittered)")
abline(lm.3b)
```



```
par(mfrow=c(2,2))
plot(lm.3b)
```



Again, this looks like a good regression model, in the sense that the assumptions underlying linear regression are approximately satisfied[1]. However, the estimated coefficient on Year, $\hat{\beta}_1 = 1.2$ with $SE(\hat{\beta}_1) = 0.73$ is not statistically significantly different from zero. From this we would conclude that there really is not enough evidence to say that there are more students repeating first grade in the 2004 school year than in 1994 school year.

(3c) Difference in the relationship of income with repeating between the two school years?

Now let's fit the interactive ANCOVA model. This will both help answer question (3c), and help us to understand whether our answers to questions (3a) and (3b) need any additional elaboration.

```
summary(lm.3c <- lm(Pct.Repeating.1st.Grade ~ Pct.Low.income.students*Year,data=houston))
```

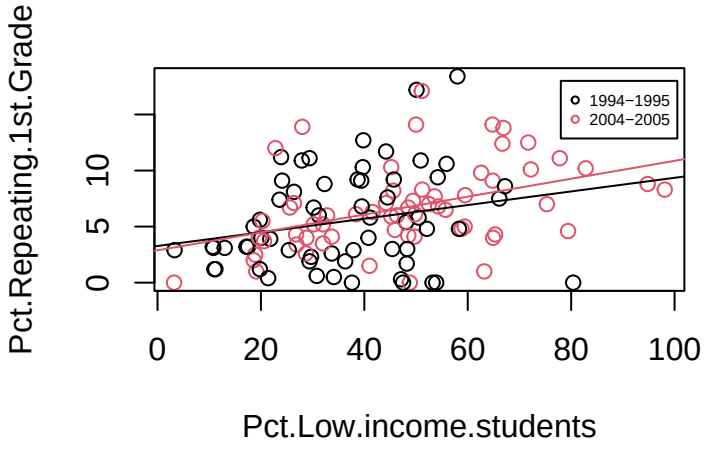
```
##
## Call:
## lm(formula = Pct.Repeating.1st.Grade ~ Pct.Low.income.students *
##     Year, data = houston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1606 -2.6121 -0.5576  1.7495 11.6014
```

[1] Remember, the clustering around different x values is just due to the fact that x is categorical. The important thing here is that we don't see any interesting patterns within or between those clusters.

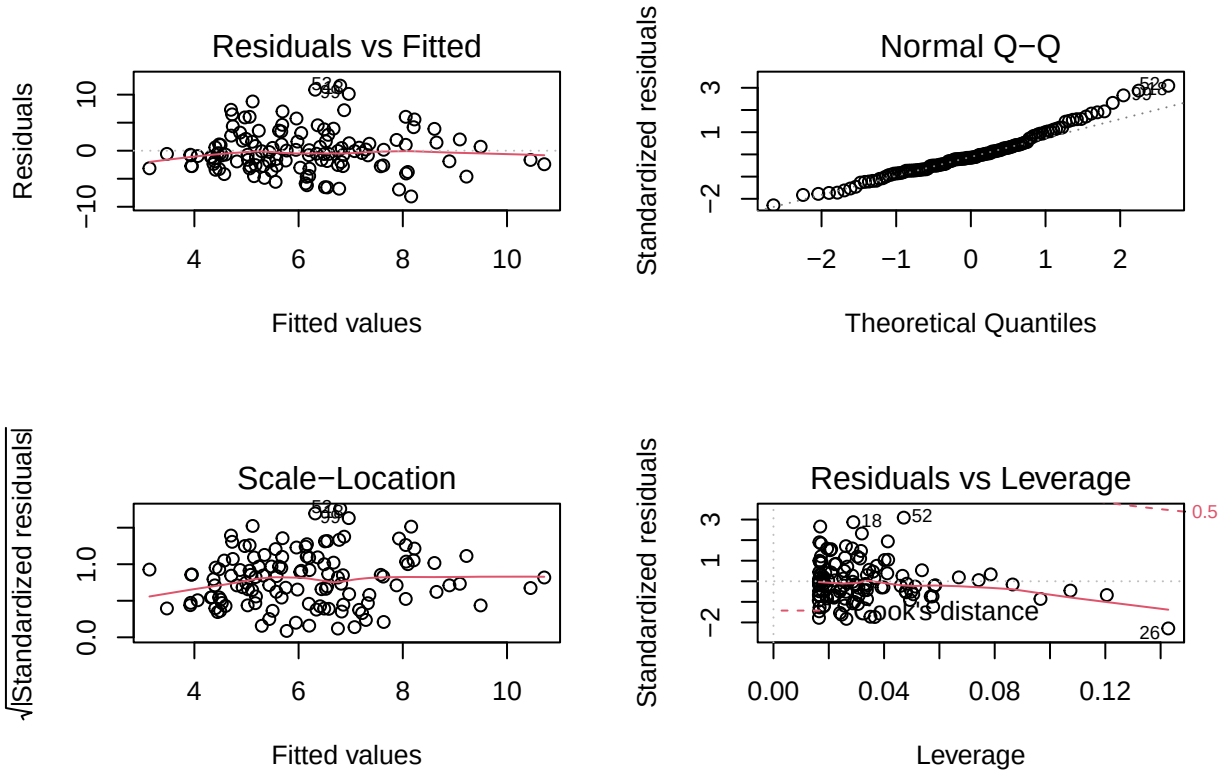
```
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       3.27194    1.22347   2.674  0.00855 **
## Pct.Low.income.students           0.06080    0.03093   1.966  0.05167 .
## Year2004                         -0.38956    1.76109  -0.221  0.82532
## Pct.Low.income.students:Year2004  0.01903    0.03949   0.482  0.63066
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.845 on 118 degrees of freedom
## Multiple R-squared:  0.1288, Adjusted R-squared:  0.1066
## F-statistic: 5.813 on 3 and 118 DF,  p-value: 0.0009689
```

The summary shows that neither the main effect for Year, nor the interaction between low income and Year, are significant in the model. This is illustrated in the scatterplot below: the regression lines for the two categories (year=1994-1995 and year=2004-2005) are nearly identical. The residual diagnostic plots that follow suggest that the modeling assumptions hold up fairly well here.

```
par(mfrow=c(1,1))
plot(Pct.Repeating.1st.Grade ~ Pct.Low.income.students,data=houston,
     col=as.numeric(houston$Year))
coefs <- coefficients(lm.3c)
abline(a=coefs[1],b=coefs[2],col=1)
abline(a=coefs[1]+coefs[3],b=coefs[2]+coefs[4],col=2)
legend(78,18,legend=c("1994-1995","2004-2005"),pch=1,col=c(1,2),cex=0.5)
```




```
par(mfrow=c(2,2))
plot(lm.3c)
```



We can also apply an F test using the ANOVA function in R to compare the full ANCOVA model in part (c) with the models in parts (a) and (b).

```
anova(lm.3a,lm.3c)
```

```
## Analysis of Variance Table
##
## Model 1: Pct.Repeating.1st.Grade ~ Pct.Low.income.students
## Model 2: Pct.Repeating.1st.Grade ~ Pct.Low.income.students * Year
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    120 1751.9
## 2    118 1744.4  2     7.512 0.2541 0.7761
```

Since the F statistic testing H_0: `Pct.Repeating.1st.Grade` $\sim$ `Pct.Low.income.students` vs H_A: `Pct.Repeating.1st.Grade` $\sim$ `Pct.Low.income.students` $*$ `Year` is nonsignificant ($F = 0.2541$, $p = 0.7761$), we cannot reject the simpler model in part (a): If we are starting with `Pct.Repeating.1st.Grade` $\sim$ `Pct.Low.income.students`, there is really no need to add `Year` (let alone an interaction with `Year`) to the model.

```
anova(lm.3b,lm.3c)
```

```
## Analysis of Variance Table
##
## Model 1: Pct.Repeating.1st.Grade ~ Year
## Model 2: Pct.Repeating.1st.Grade ~ Pct.Low.income.students * Year
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    120 1957.9
## 2    118 1744.4  2    213.52 7.2221 0.001099 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here the F statistic testing H_0: `Pct.Repeating.1st.Grade` $\sim$ `Year` vs H_A: `Pct.Repeating.1st.Grade` $\sim$ `Pct.Low.income.students` $*$ `Year` is highly significant ($F = 7.2221$, $p = 0.001099$), so we would reject the simpler model in part (b): starting with `Pct.Repeating.1st.Grade` $\sim$ `Year`, we really do get a better model by adding `Pct.Low.income.students` to the model.

Note that we cannot perform an F test like `anova(lm.3a, lm.3b)` directly[2], since these models are not nested. We can infer, informally, from the above tests though, that of all three models, `Pct.Repeating.1st.Grade` $\sim$ `Pct.Low.income.students` seems to be the best. Thus, there is no difference between years in the way that low income is related to repeating first grade.

4. (course evaluations question from Gelman & Hill)

(4a)

```
beauty <- read.csv("ProfEvaltnsBeautyPublic.csv")
model1 <- lm(courseevaluation ~ btystdave + female, data = beauty)
model2 <- lm(courseevaluation ~ btystdave*female, data = beauty)
coef1 <- coef(model1)  # Get model parameters
coef2 <- coef(model2)  # Get model parameters
```

(i)

```
par(mfrow = c(1, 2), oma = c(2, 0, 1, 0), mar = c(4, 4, 1, 2))
# First model
plot(courseevaluation ~ btystdave, data = beauty)
abline(a = coef1["(Intercept)"], b = coef1["btystdave"], col = "red")  # line for males
abline(a = coef1["(Intercept)"] + coef1["female"],
       b = coef1["btystdave"], col = "blue")  # line for females

# Second model
plot(courseevaluation ~ btystdave, data = beauty)
abline(a = coef2["(Intercept)"], b = coef2["btystdave"], col = "red")  # line for males
abline(a = coef2["(Intercept)"] + coef2["female"],
       b = coef2["btystdave"] + coef2["btystdave:female"], col = "blue")  # line for females

legend(0.5, 2.6, legend = c("Males", "Females"),
       col = c("red", "blue"), lty = c(1, 1), cex = 0.6)
```

The red lines represent the predicted values for males, and the blue lines represent predicted values for females. In the first plot, the intercept for the red line is the `(Intercept)` coefficient, while the intercept for the blue line is the `(Intercept)` coefficient plus the `female` coefficient. The slope for both lines is the `btystdave` coefficient.

In the second plot, the intercept for the red line is the `(Intercept)` coefficient. The intercept for the blue line again is the `(Intercept)` coefficient plus the `female` coefficient. The slope for the red line is the `btystdave` coefficient, while the slope for the blue line is the `btystdave` coefficient plus the `btystdave:female` coefficient.

[2]There are ways to compare non-nested models, e.g. cross-validation, AIC, BIC, etc., but it doesn't work with F tests or likelihood ratio tests.

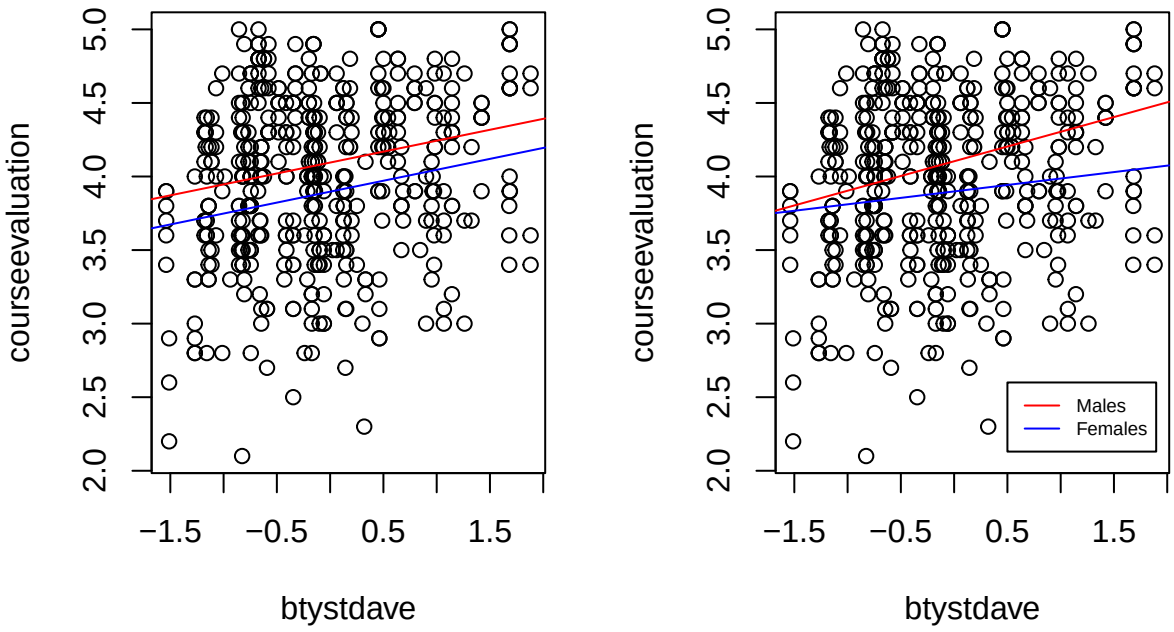


Figure 1: Problem 2(a)i: Course evaluations against beauty ratings, with no interaction (left) and with an interaction (right).

(ii)

```
par(mfrow=c(2, 2), mar = c(4, 4.5, 2, 1))
plot(model1)
```

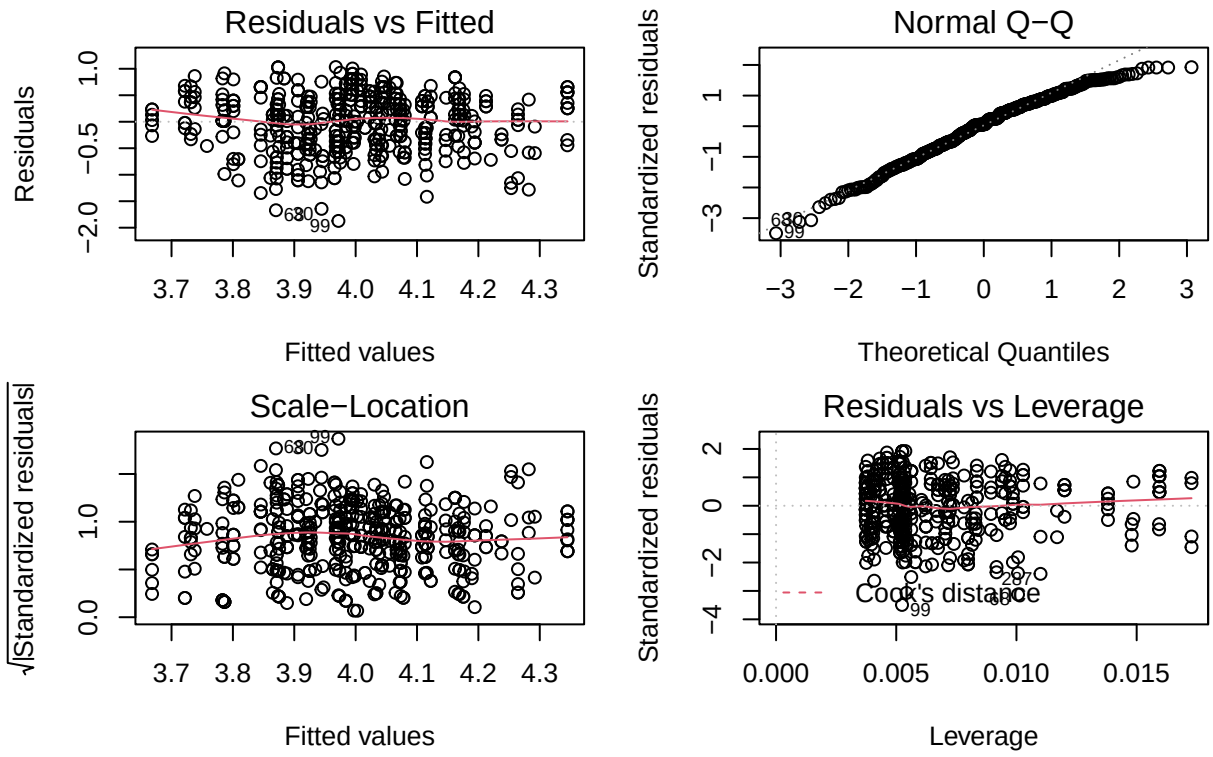


Figure 2: Problem 2(a)ii: Diagnostic plots for the first model

In Figure 2, there don't appear to be any worrying trends between the fitted values and the residuals. However, the Q-Q plot shows some non-normality in the residuals, on the right side of the plot; and the bottom right plot appears to show that the residuals are left skewed, as there are a good number of standardized residuals with values below -2. This is apparent also looking back at the scatter plots in Figure 1.

```
par(mfrow=c(2, 2), mar = c(4, 4.5, 2, 1))
plot(model2)
```

Figure 3 for the second model, with the interaction effect, shows essentially the same results as for the first model. It's possible that the fit could be improved by transforming the predictor(s) and/or the outcome.

(iii)

```
summary(model1)
```

Here's the summary for the first model with no interaction:

```
##
## Call:
## lm(formula = courseevaluation ~ btystdave + female, data = beauty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

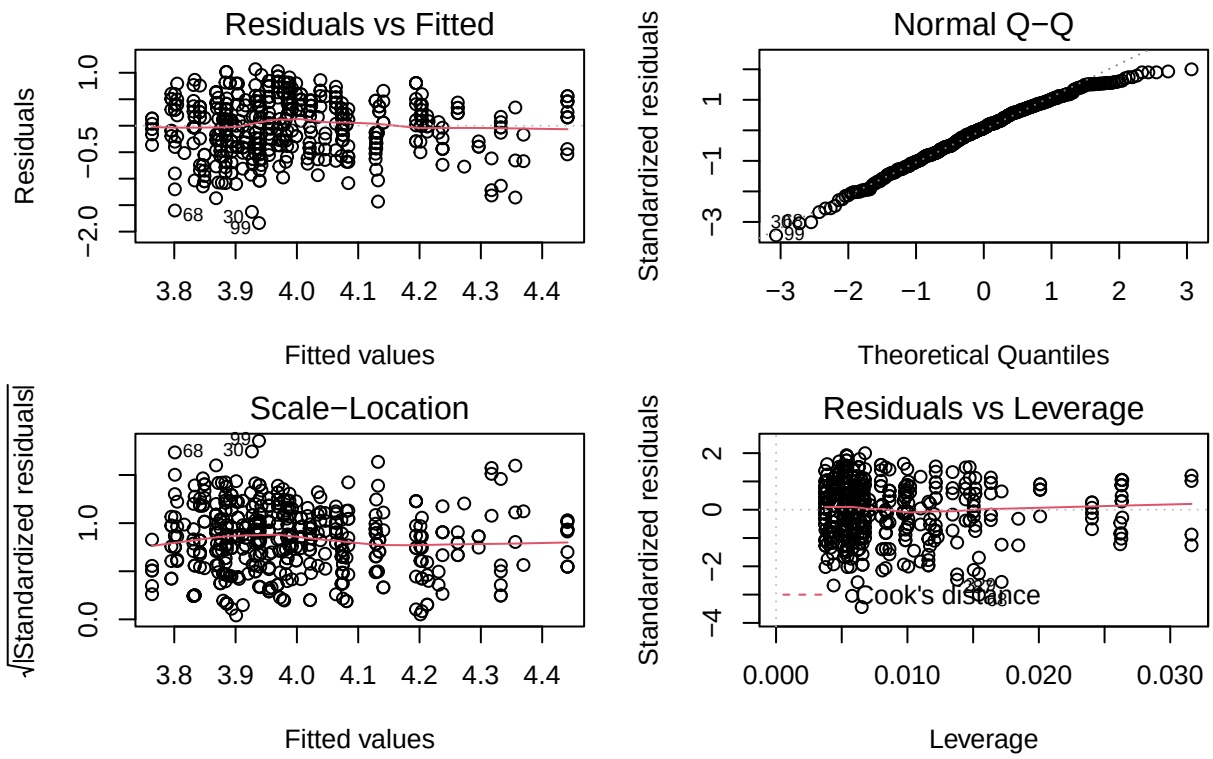


Figure 3: Problem 2(a)ii: Diagnostic plots for the second model

```
## -1.87196 -0.36913  0.03493  0.39919  1.03237
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.09471    0.03328  123.03  < 2e-16 ***
## btystdave    0.14859    0.03195    4.65 4.34e-06 ***
## female      -0.19781    0.05098   -3.88  0.00012 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5373 on 460 degrees of freedom
## Multiple R-squared:  0.0663, Adjusted R-squared:  0.06224
## F-statistic: 16.33 on 2 and 460 DF,  p-value: 1.407e-07
```

```
summary(model2)
```

Here's the summary for the second model, which has the interaction:

```
##
## Call:
## lm(formula = courseevaluation ~ btystdave * female, data = beauty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83820 -0.37387  0.04551  0.39876  1.06764
##
```

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        4.10364    0.03359 122.158  < 2e-16 ***
## btystdave          0.20027    0.04333   4.622 4.95e-06 ***
## female            -0.20505    0.05103  -4.018 6.85e-05 ***
## btystdave:female -0.11266    0.06398  -1.761   0.0789 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5361 on 459 degrees of freedom
## Multiple R-squared:  0.07256,    Adjusted R-squared:  0.0665
## F-statistic: 11.97 on 3 and 459 DF,  p-value: 1.471e-07
```

```
anova(model1, model2)
```

Here's a partial F-test, computed with `anova(model1, model2)`, where `model1` is the model with no interaction:

```
## Analysis of Variance Table
##
## Model 1: courseevaluation ~ btystdave + female
## Model 2: courseevaluation ~ btystdave * female
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1    460 132.81
## 2    459 131.92  1   0.89124 3.101 0.07891 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficient estimates for `btystdave` and `female` are quite large in both models with respect to their estimated standard errors, which is reflected in the fact that their associated p-values are quite small: the three stars (***) mark that they are less than 0.001. Notice that the estimates differ between the two models. This is not surprising; in general, adding a new variable to a model will change the estimates of all the coefficients, unless that new variable is completely uncorrelated with the previous variables. Since the "new" variable here is an interaction between the other two variables, it is naturally correlated with those variables, so we should expect the estimates to change.

The adjusted R^2 values for both models are quite small, indicating that the predictors do not account for much of the total variance. The addition of the interaction term did not increase the adjusted R^2 by a non-trivial amount.

The estimate of `btystdave:female`, the interaction term, is not significantly different from 0 at the conventional $\alpha = 0.05$ level. The actual magnitude of the estimate is also smaller than the main effect estimates. The partial F-test here constitutes a test of the null hypothesis

$$H_0 : \beta_3 = 0$$

where β_3 is the coefficient associated with `btystdave:female`. Hence, the p-value of the partial F-test that is printed in the `anova` output is equivalent to the p-value of the t-test in the output of `summary(model2)`, namely, $p = 0.0789$.

Since we don't have enough evidence to reject H_0, and since the addition of the interaction term term did not increase the adjusted R^2, in the interest of parsimony, we may decide based on these results to exclude the interaction term.

(4b)

We don't have a lot of formal tools available for variable selection at this stage, so the general approach here is to think in real-world terms about what variables seem likely to be related to the outcome and see what happens when they are included in the model. There's no one correct approach or outcome here. In addition to `btystdave` and `female`, let's consider the variables `age`, `minority`, `nonenglish`, `tenured`, and `onecredit`. First, let's look at a pairs plot with `courseevaluation`:

```
cols <- c("minority", "nonenglish", "tenured", "onecredit")
beauty[, cols] <- lapply(beauty[, cols], as.factor)
ggpairs(beauty[, c("courseevaluation", "age", cols)])
```

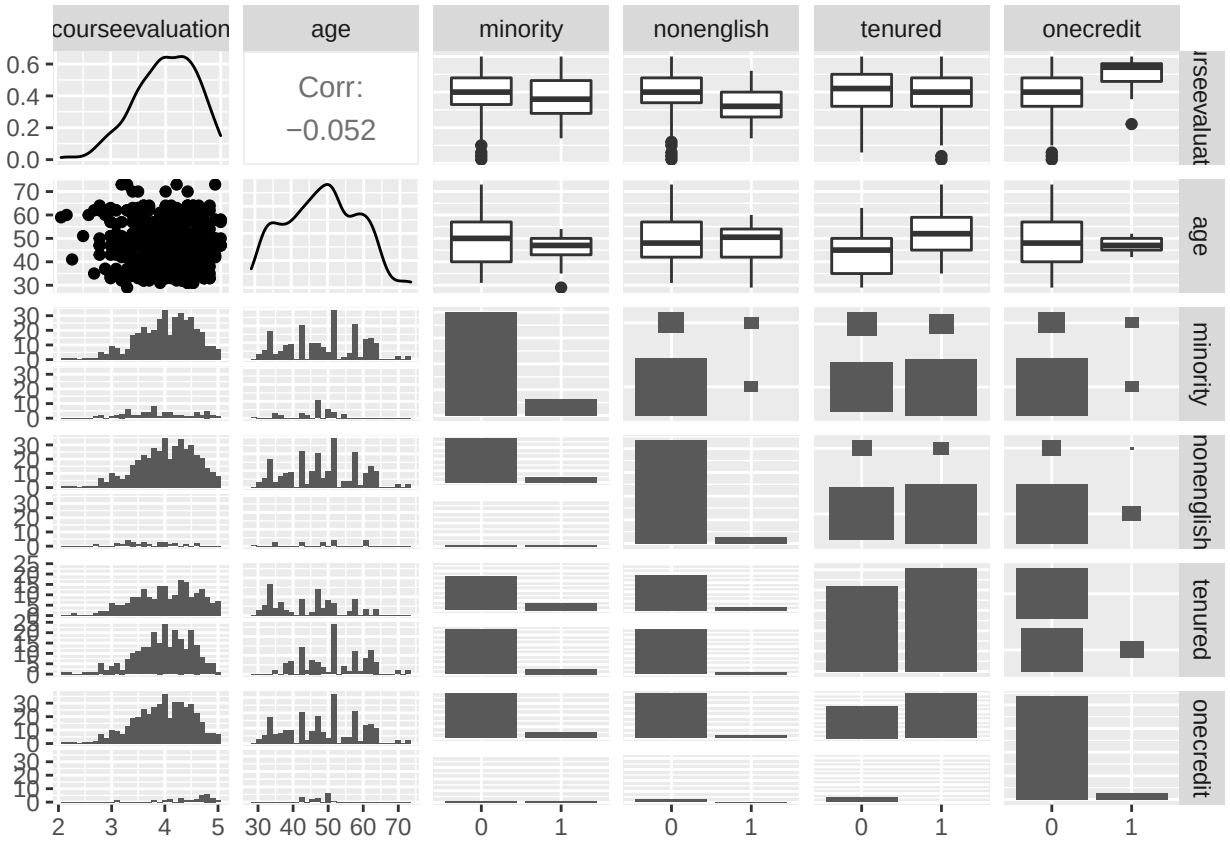


Figure 4: Pairs plot for candidate predictors in 2(b)

The marginal distributions of each variable appear on the diagonal, while the top row and leftmost column give different visualizations of the relationship between the course evaluation scores and the other five variables. Among the candidate predictors, only `onecredit` shows a clear marginal relationship with `courseevaluation`, but since these are marginal plots, that doesn't mean the other variables won't be informative to some degree. There are very few one credit courses represented here (27 out of 463 rows), but nevertheless, let's try including it, along with the following sets of variables:

Model m1: onecredit

Model m2: onecredit, minority, nonenglish

Model m3: onecredit, minority, nonenglish, tenured, age

For each model, we'll look at a model summary and diagnostic plots. We can also run an F-test comparing the model from part 2(a)ii (aka `model1`) to the expanded model.

Model m1

```
m1 <- lm(courseevaluation ~ btystdave + female + onecredit)
```

```
m1 <- lm(courseevaluation ~ btystdave + female + onecredit, data = beauty)
summary(m1)
```

```
##
## Call:
## lm(formula = courseevaluation ~ btystdave + female + onecredit,
##     data = beauty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82352 -0.34541  0.06084  0.38657  1.08122
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.05783    0.03287 123.438  < 2e-16 ***
## btystdave    0.16258    0.03103   5.240 2.45e-07 ***
## female      -0.18832    0.04938  -3.814 0.000155 ***
## onecredit1   0.58513    0.10358   5.649 2.84e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5201 on 459 degrees of freedom
## Multiple R-squared:  0.127,  Adjusted R-squared:  0.1213
## F-statistic: 22.25 on 3 and 459 DF,  p-value: 1.804e-13
```

```
par(mfrow=c(2, 2), mar = c(4, 4.5, 2, 1))
plot(m1)
```

```
anova(model1, m1)
```

F-Test for model m1 vs. model1

```
## Analysis of Variance Table
##
## Model 1: courseevaluation ~ btystdave + female
## Model 2: courseevaluation ~ btystdave + female + onecredit
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    460 132.81
## 2    459 124.18  1    8.6326 31.909 2.838e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model m2

```
m2 <- lm(courseevaluation ~ btystdave + female + onecredit + minority + nonenglish)
```

```
m2 <- lm(courseevaluation ~ btystdave + female + onecredit + minority + nonenglish,
         data = beauty)
summary(m2)
```

```
##
## Call:
```

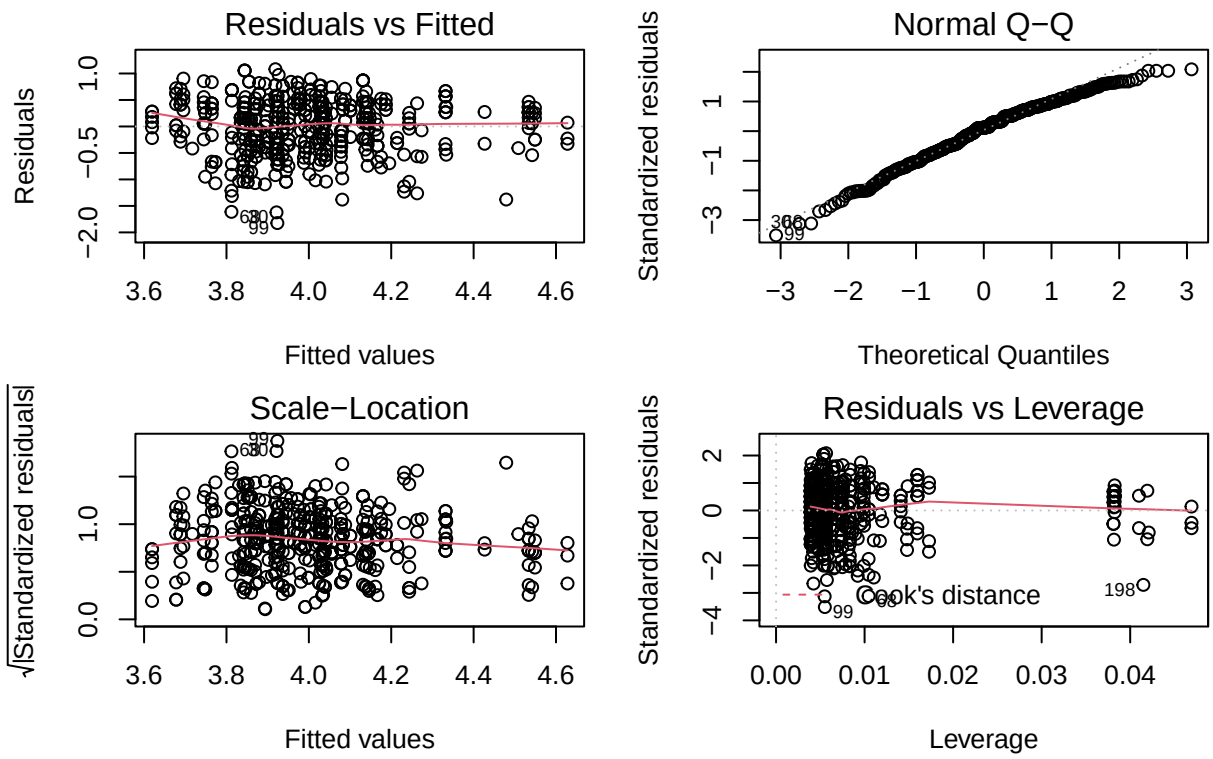


Figure 5: Diagnostic plots for model m1

```
## lm(formula = courseevaluation ~ btystdave + female + onecredit +
##     minority + nonenglish, data = beauty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.84951 -0.33198  0.04644  0.37907  1.05533
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.08668    0.03327 122.833  < 2e-16 ***
## btystdave    0.16604    0.03063   5.422 9.59e-08 ***
## female      -0.17418    0.04911  -3.546 0.000431 ***
## onecredit1   0.64133    0.10632   6.032 3.34e-09 ***
## minority1   -0.16479    0.07569  -2.177 0.029982 *
## nonenglish1 -0.24801    0.10523  -2.357 0.018859 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.513 on 457 degrees of freedom
## Multiple R-squared:  0.1546, Adjusted R-squared:  0.1453
## F-statistic: 16.71 on 5 and 457 DF,  p-value: 3.586e-15
```

```
par(mfrow=c(2, 2), mar = c(4, 4.5, 2, 1))
plot(m2)
```

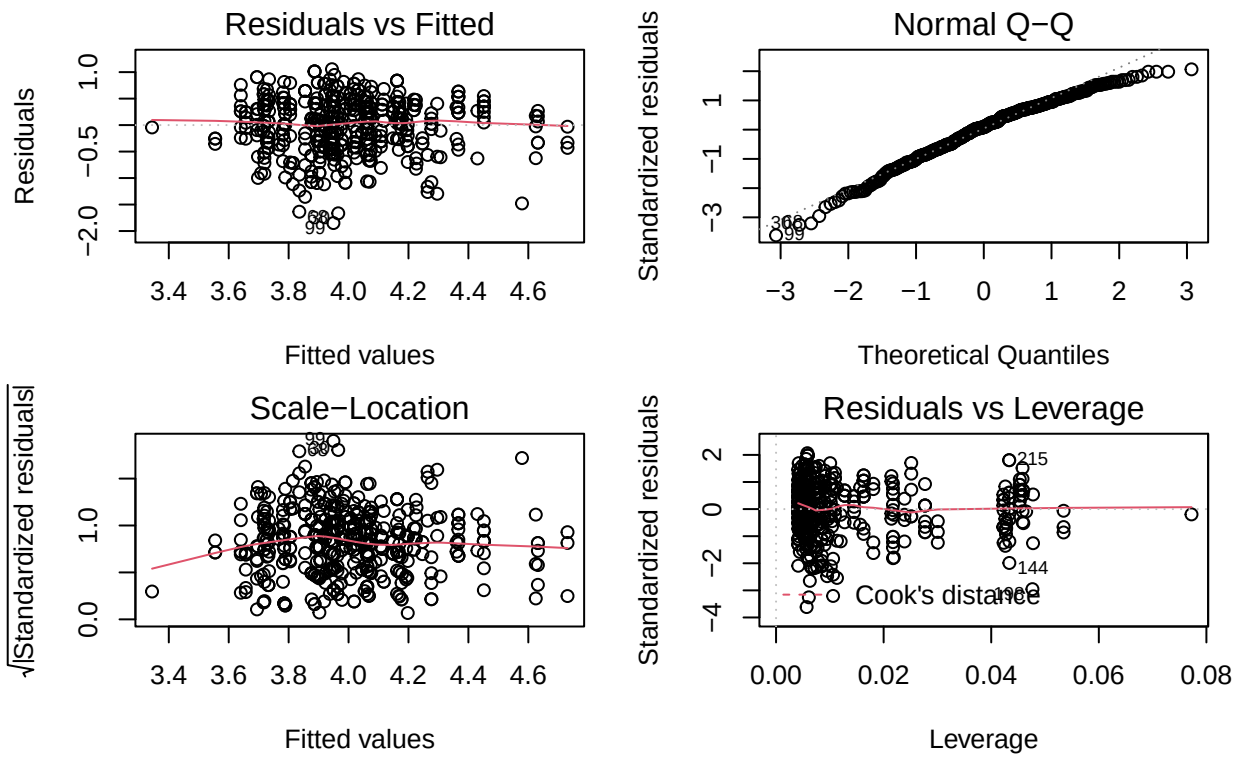


Figure 6: Diagnostic plots for model m2

```
anova(model1, m2)
```

F-Test for model m2 vs. model1

```
## Analysis of Variance Table
##
## Model 1: courseevaluation ~ btystdave + female
## Model 2: courseevaluation ~ btystdave + female + onecredit + minority +
##     nonenglish
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    460 132.81
## 2    457 120.25  3    12.556 15.906 7.475e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model m3

```
m3 <- lm(courseevaluation ~ btystdave + female + onecredit + minority + nonenglish)
```

```
m3 <- lm(courseevaluation ~ btystdave + female + onecredit + minority + nonenglish +
           tenured + age, data = beauty)
summary(m3)
```

```
##
## Call:
## lm(formula = courseevaluation ~ btystdave + female + onecredit +
##     minority + nonenglish + tenured + age, data = beauty)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8386 -0.3414  0.0501  0.3776  1.0619
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.187474   0.138209  30.298  < 2e-16 ***
## btystdave    0.159636   0.032038   4.983 8.92e-07 ***
## female      -0.182937   0.051959  -3.521 0.000474 ***
## onecredit1   0.640845   0.112170   5.713 2.01e-08 ***
## minority1   -0.168522   0.076030  -2.217 0.027151 *
## nonenglish1 -0.245503   0.106022  -2.316 0.021024 *
## tenured1     0.005034   0.056134   0.090 0.928583
## age         -0.002068   0.002807  -0.737 0.461551
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5138 on 455 degrees of freedom
## Multiple R-squared:  0.1556, Adjusted R-squared:  0.1426
## F-statistic: 11.98 on 7 and 455 DF,  p-value: 4.703e-14
```

```
par(mfrow=c(2, 2), mar = c(4, 4.5, 2, 1))
plot(m3)
```

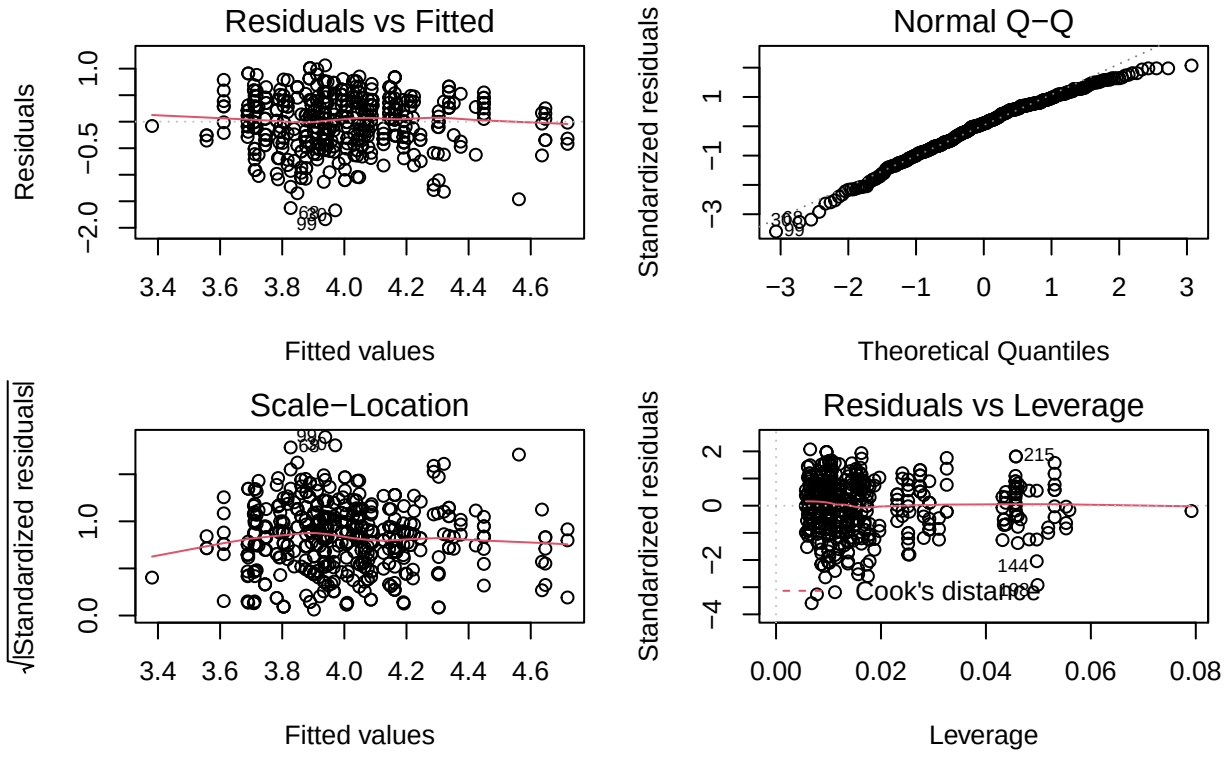


Figure 7: Diagnostic plots for model m3

```
anova(model1, m3)
```

F-Test for model m3 vs. model1

```
## Analysis of Variance Table
##
## Model 1: courseevaluation ~ btystdave + female
## Model 2: courseevaluation ~ btystdave + female + onecredit + minority +
##     nonenglish + tenured + age
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1    460 132.81
## 2    455 120.10  5    12.706 9.627 9.498e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The diagnostics for all three models look reasonable, and they look fairly similar to the diagnostics for `model1`, the model that we retained from part 2(a)iii. The F-tests comparing each model to the baseline `model1` are all highly significant. Summaries for models `m1` and `m2` each model show highly significant coefficients. In model `m3`, however, the tests for `tenured` and `age` are non-significant, and the coefficient estimates are tiny. Given these tiny estimates, it would be reasonable to drop these variables from the model and retain model `m2`.

If we believe this model, then we can say that course evaluations bear at least a roughly linear relationship to ratings of beauty, whether a person is female or male, whether a course is one credit or not, whether the professor is a minority, and whether the professor is a native English speaker or not, with the sign and magnitude of the estimated coefficients indicating the nature and strength of the relationships. It is tempting to draw causal conclusions here, but we can't really do that without checking some additional assumptions which are beyond the scope of the problem. However, these patterns may at least be suggestive of causal relationships, which we can subsequently go about testing more rigorously.

(end of solutions)