Homework 03 Solutions

9/24/2018

1.

(1a)

We will utilize 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

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

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:

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

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

(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:

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

$\square$

(iii)

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

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

□

(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:

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

□

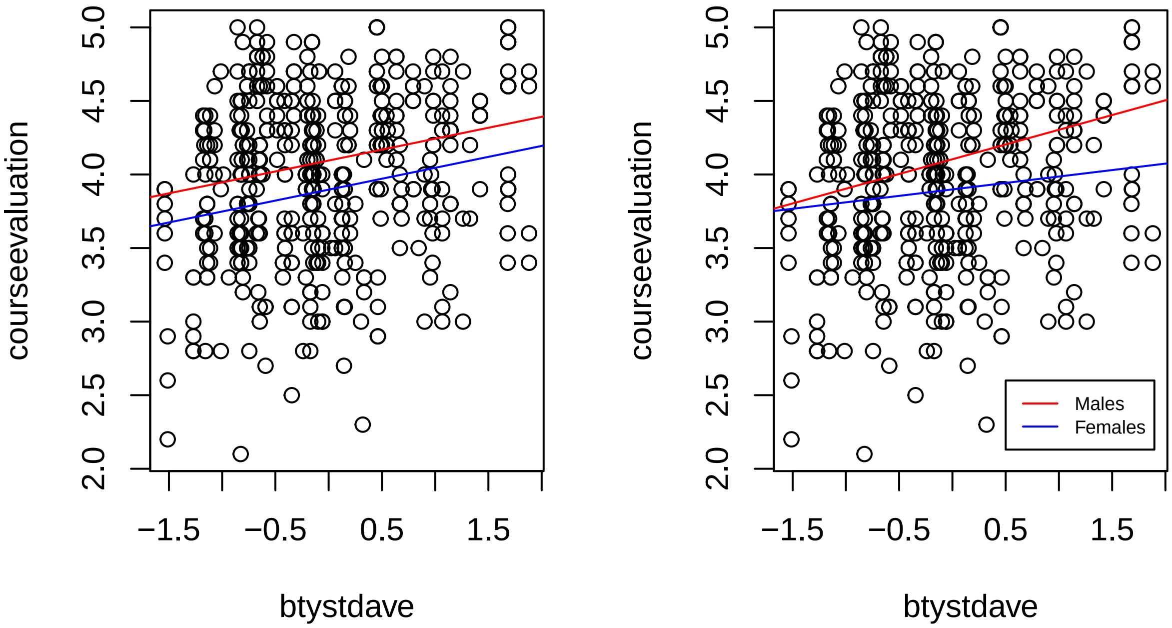


Figure 1: Problem 2(a)i: Course evaluations against beauty ratings, with no interaction (left) and with an interaction (right).

2.

(2a)

(i)

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.

(ii)

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.

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)

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

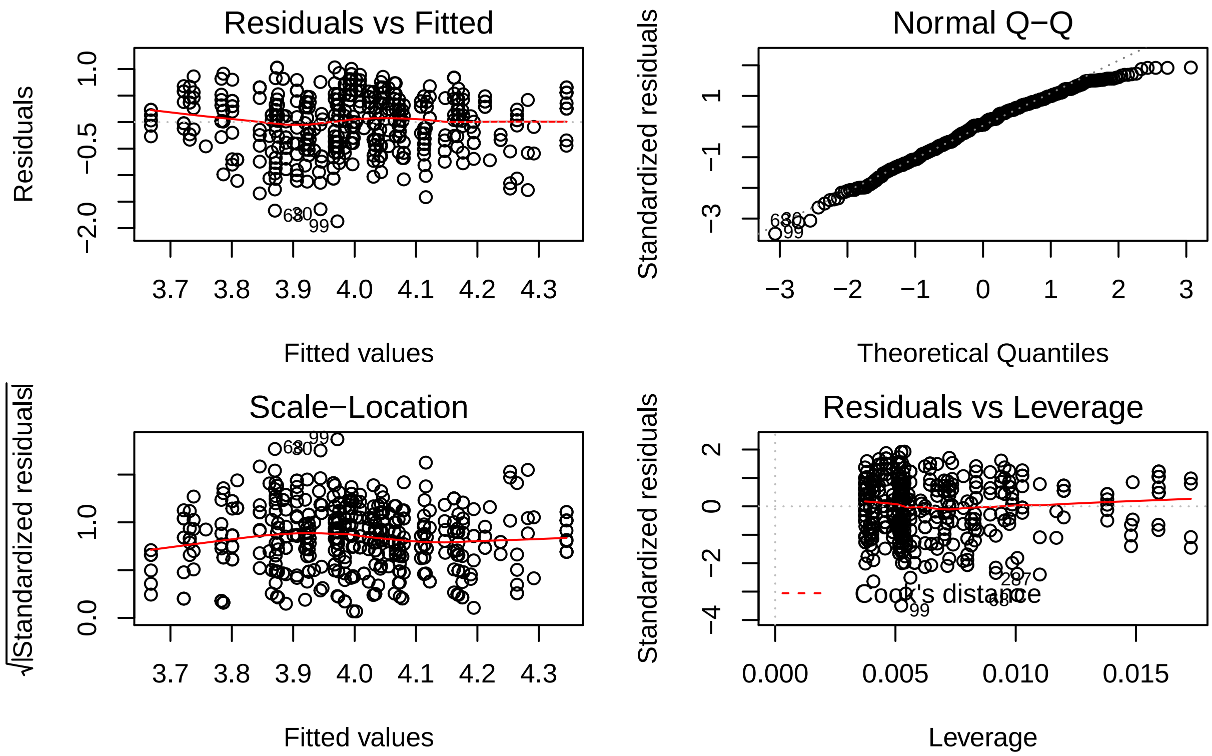


Figure 2: Problem 2(a)ii: Diagnostic plots for the first model

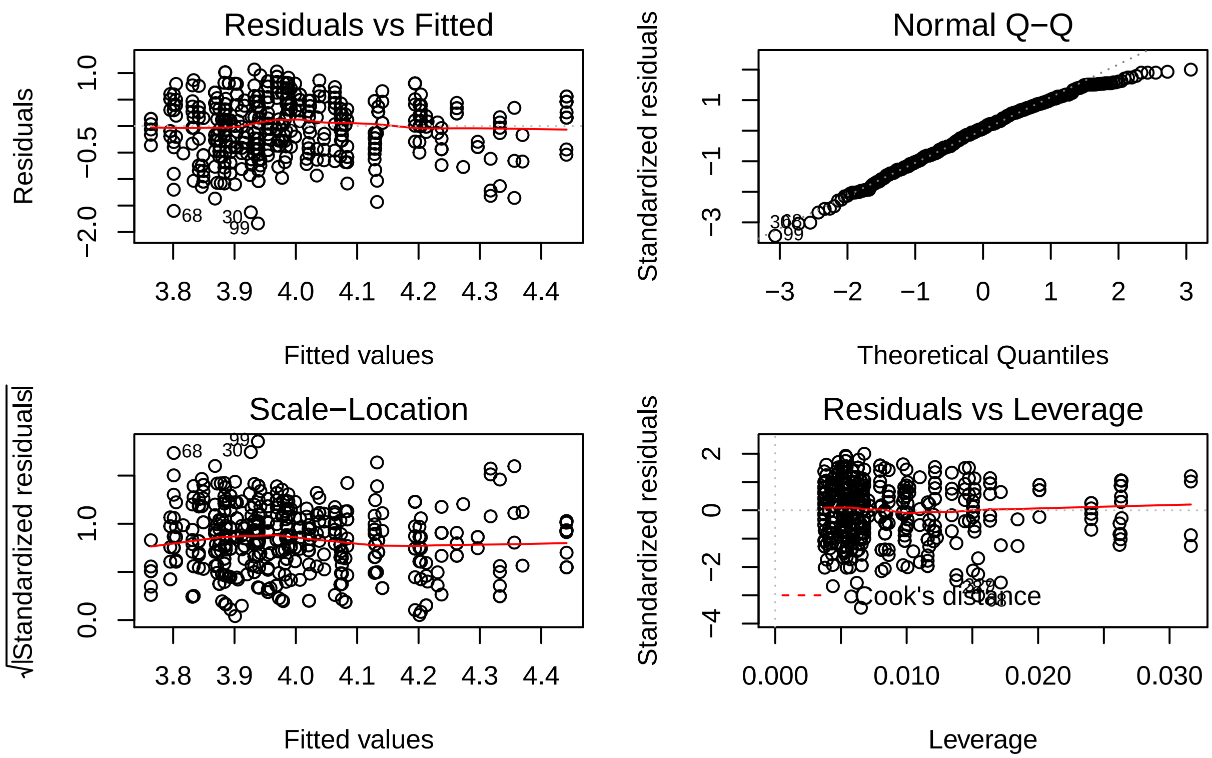


Figure 3: Problem 2(a)ii: Diagnostic plots for the second model

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

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
```

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

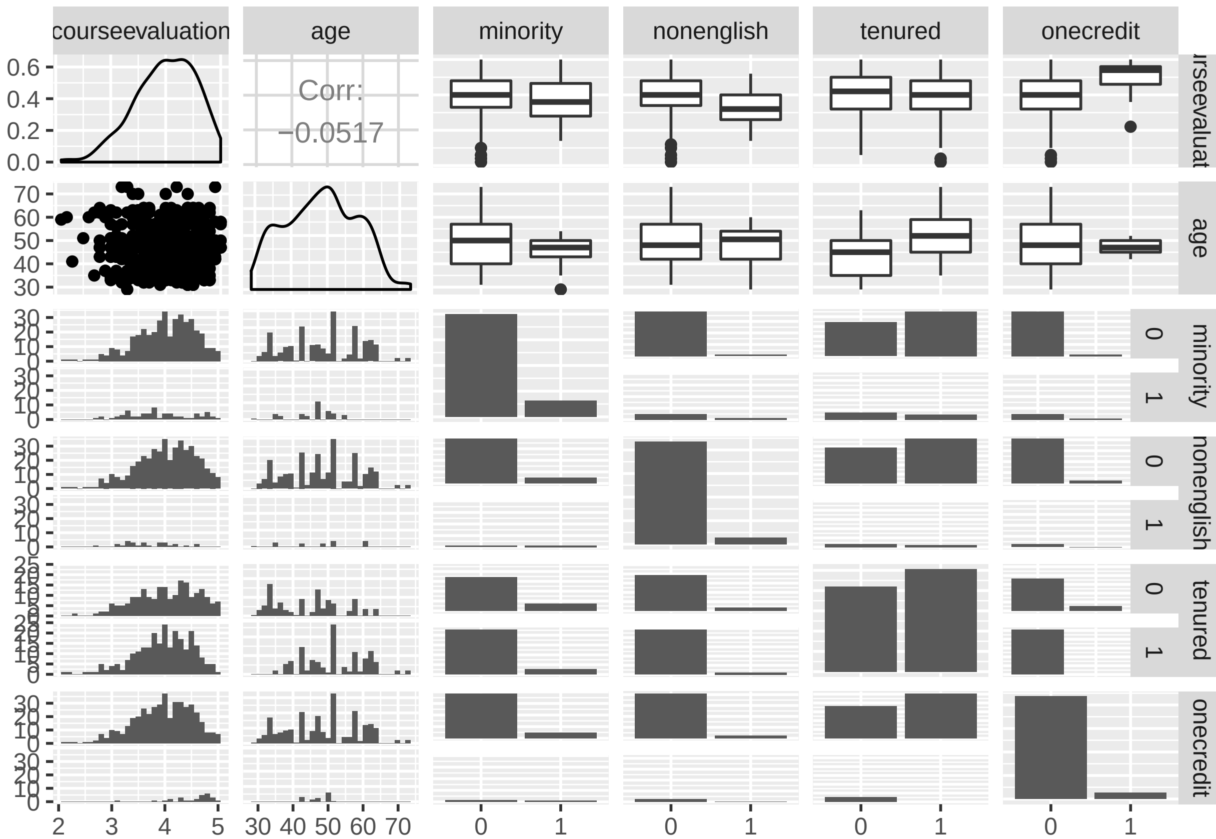
$$H_0 : \beta_3 = 0$$

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.

(2b)

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`:



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)
```

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

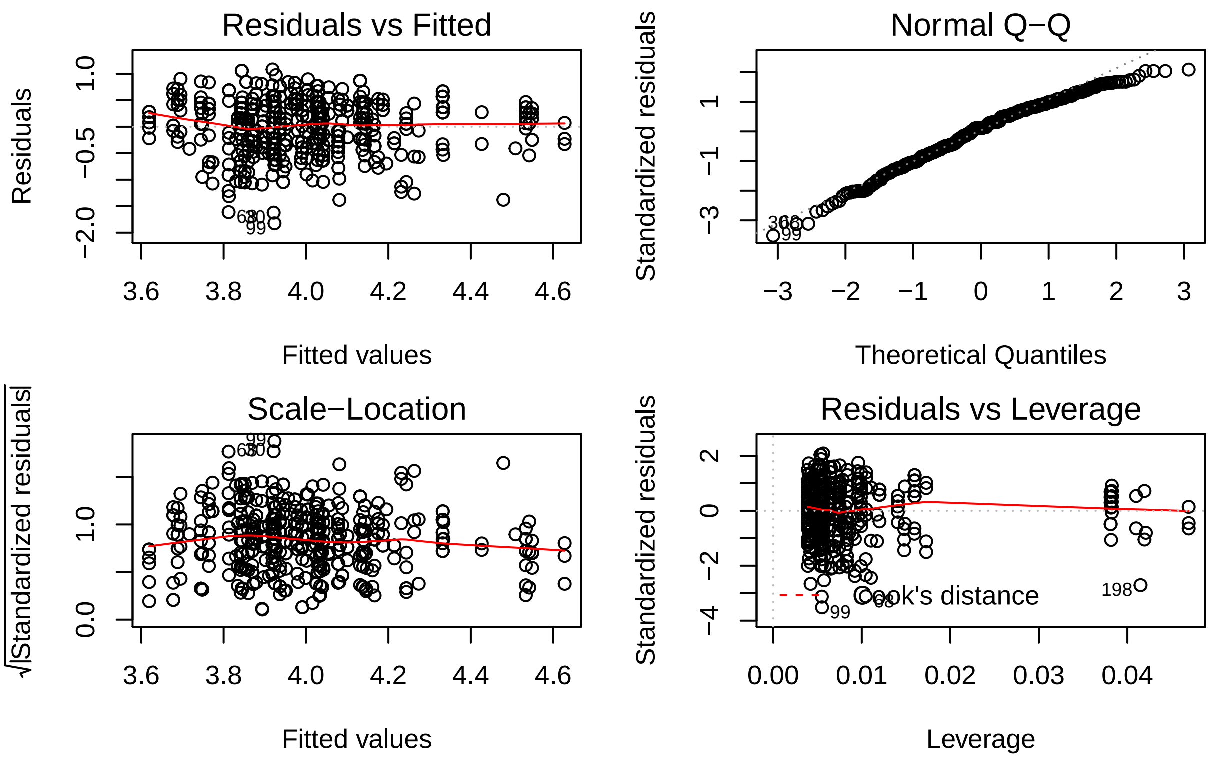


Figure 4: Diagnostic plots for model m1

```
## -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
```

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
```

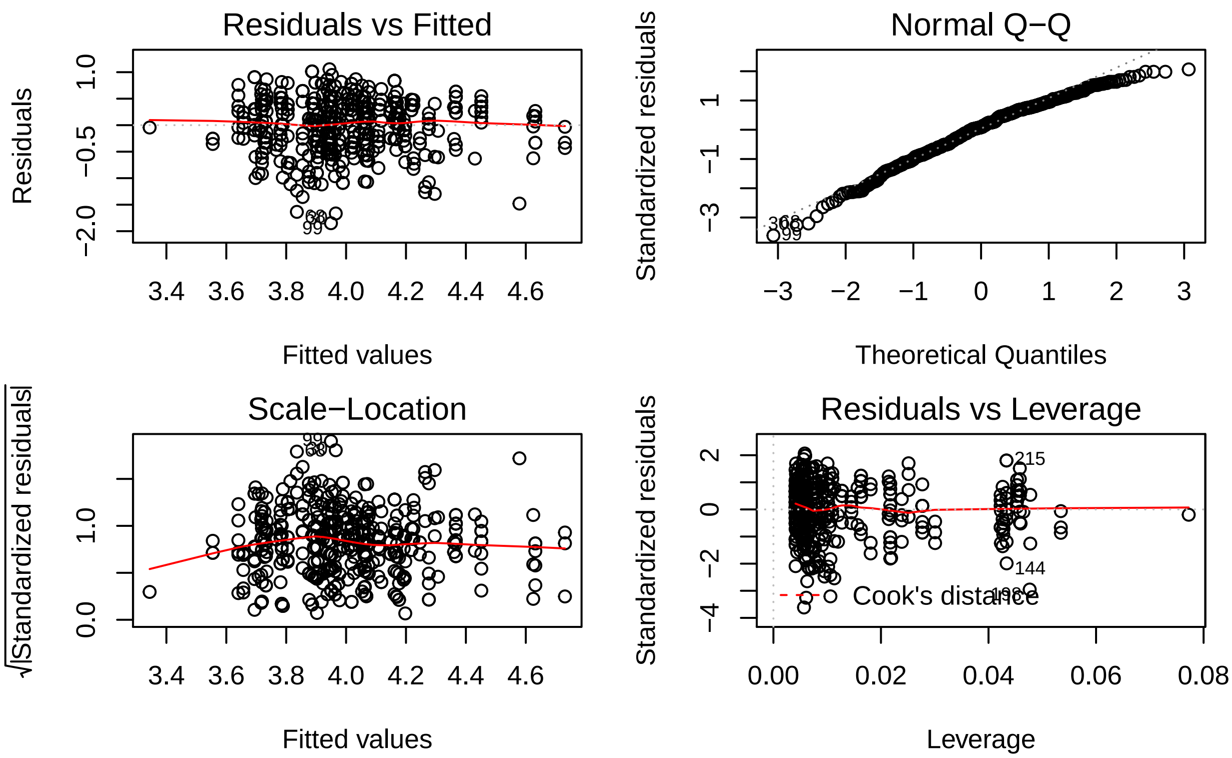


Figure 5: Diagnostic plots for model m2

Model m2

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

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

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)
```

```
##
## 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
```

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 ***
## ---
```

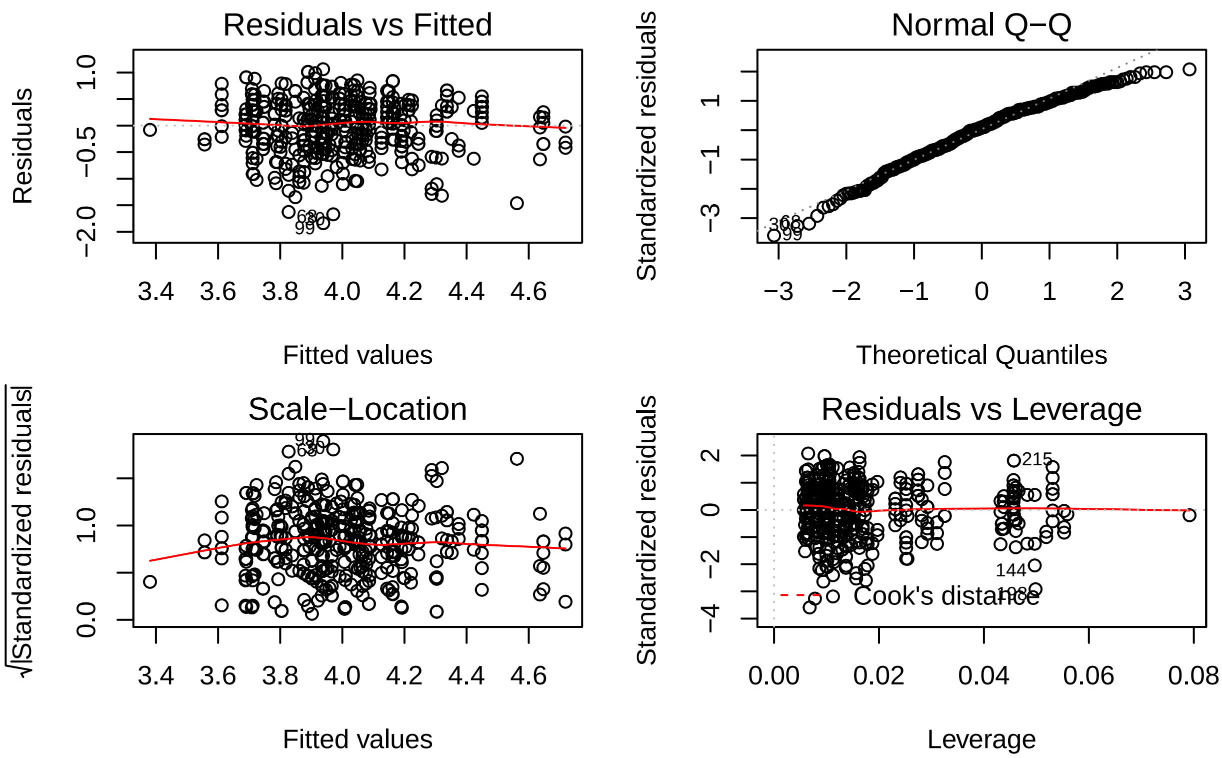


Figure 6: Diagnostic plots for model m3

```
## 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.

3. (Sheather 5.2)

(3a)

This question is ambiguous. Here, I'll interpret it to refer to a change across years, so I'll plot the change in percentage of students repeating first grade against the percentage of low income students, along with the least squares line:

This line has a very slight negative slope, so an increase in the percentage of students repeating first grade does not appear to be associated with an increase in the percentage of low income students. Note that

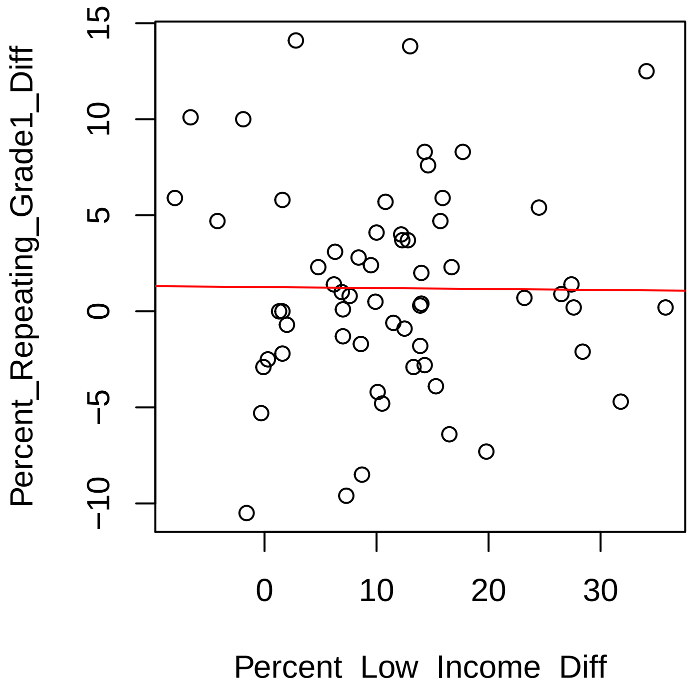


Figure 7: Change in percent repeating grade 1 vs. percent low income, 2004 - 1994

this is not what is typically referred to as an analysis of covariance, since that usually means that we are interested in the effect of a categorical predictor in the presence of a continuous predictor. Here, there's only a continuous predictor.

(3b)

Let's plot the percentage repeating grade 1 against the year, and let's perform the corresponding regression:

```
##
## Call:
## lm(formula = Percent_Repeating_Grade1 ~ Year, data = reading)
##
## 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
```

There is no evidence of a difference between the years. Again, this is not what is typically called an analysis of covariance, but there's no need to include additional predictors to answer the question.

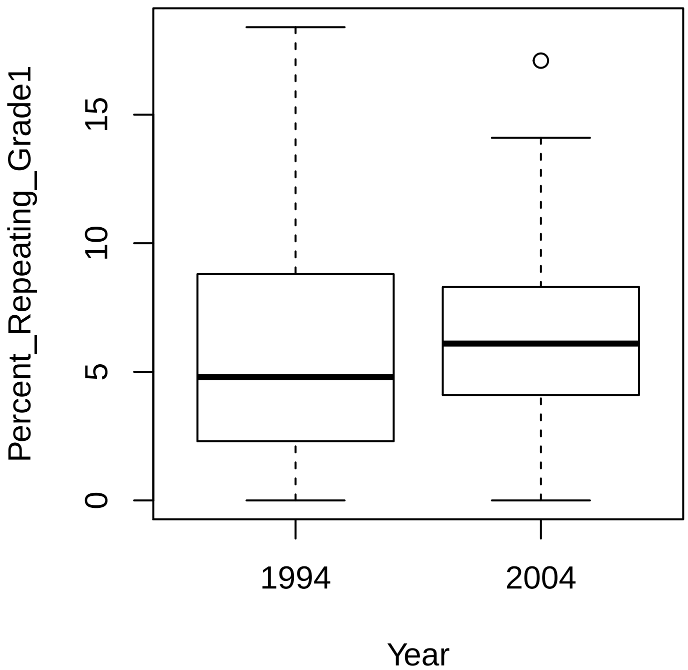


Figure 8: Percent repeating grade 1 vs. year

(3c)

Let's compare two models, one with effects of `Year` and `Percent_Low_Income` but no interaction between them, and one with an interaction. We'll use `anova()` to compare the models.

```
## Analysis of Variance Table
##
## Model 1: Percent_Repeating_Grade1 ~ Year * Percent_Low_Income
## Model 2: Percent_Repeating_Grade1 ~ Year + Percent_Low_Income
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    118 1744.4
## 2    119 1747.8 -1   -3.4351 0.2324 0.6307
```

There's insufficient evidence to reject the null hypothesis that the coefficient for the interaction term is 0. In other words, there's no evidence that the relationship between the percentage of students repeating first grade and the percentage of low income students differs across years.