Homework 2

Due Date

March 2nd at 4pm

Tricky Questions

State whether you agree or disagree with the following statements, and explain your reasoning.

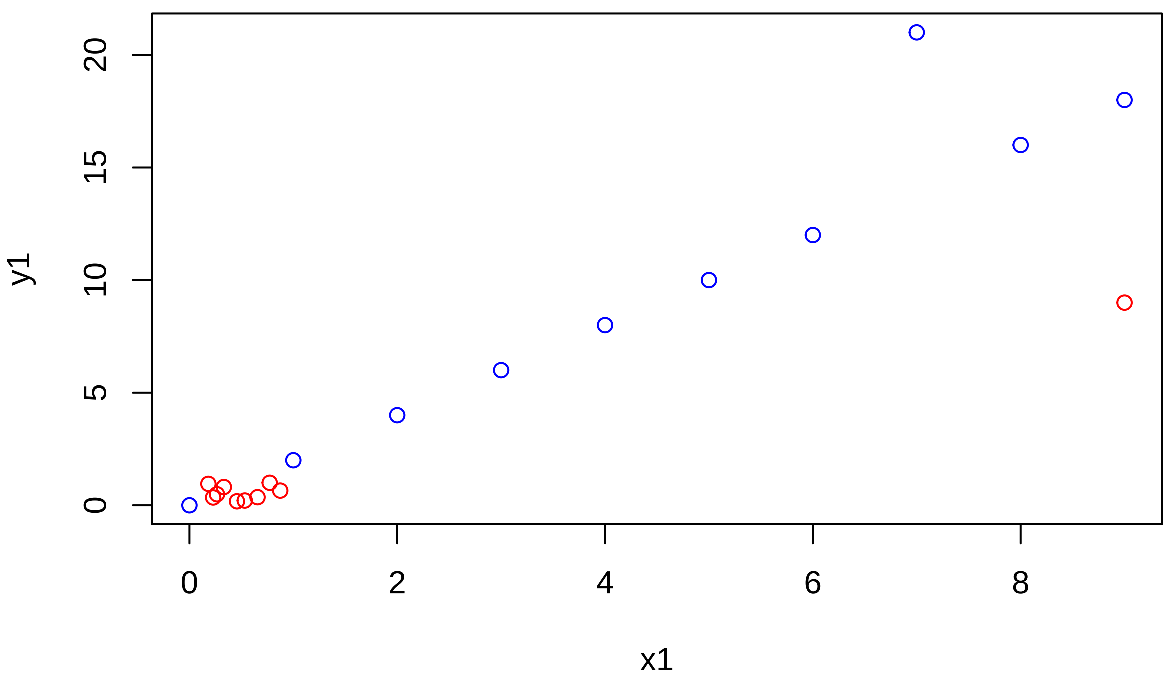
a. Removing an outlier or high leverage point always increases R^2. **False**. *It depends on the data! Examples:*

```
library(tidyverse)
library(magrittr)

x1 = seq(0,9,1)
y1 = 2*x1
y1[8] = 3*x1[8]

x2 = runif(10)
y2 = runif(10)
x2[10] = 9
y2[10] = 9


plot(x1,y1,col='blue')
points(x2,y2,col='red')
```



```r
print('Example 1, outlier:')
```

```
## [1] "Example 1, outlier:"
```

```r
summary(lm(y1~x1))$r.squared
```

```
## [1] 0.9090568
```

```r
print('Example 1, no outlier:')
```

```
## [1] "Example 1, no outlier:"
```

```r
summary(lm(y1[-8]~x1[-8]))$r.squared
```

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

```r
print('Example 2, outlier:')
```

```
## [1] "Example 2, outlier:"
```

```r
summary(lm(y2~x2))$r.squared
```

```
## [1] 0.9816423
```

```r
print('Example 2, no outlier:')
```

```
## [1] "Example 2, no outlier:"
```

```r
summary(lm(y2[-10]~x2[-10]))$r.squared
```

```
## [1] 0.00314764
```

b. If the correlation matrix between all independent variables in a regression model has an off-diagonal element near 1, then that indicates that at least one pair of independent variables are *collinear* with each other **True** *This is just the definition of collinearity.*

c. The numerical values chosen for a dummy variable do not impact the performance of the regression model **True-ish** *The numerical values themselves don't matter, but the spacing does!*

RABE 3.3

A teacher has created a dataset containing the scores on a final examination F, as well as the scores in two preliminary examinations P_1 and P_2 for 22 students in a statistics course. The data can be found on Canvas under `Files>data>exams.csv`.

a. Fit each of the following models to the data:

$$\begin{aligned}
\text{Model 1: } & F_i = \beta_0 + \beta_1 P_{1i} + \epsilon_i \\
\text{Model 2: } & F_i = \beta_0 + \beta_2 P_{2i} + \epsilon_i \\
\text{Model 3: } & F_i = \beta_0 + \beta_1 P_{1i} + \beta_2 P_{2i} + \epsilon_i
\end{aligned}$$

b. Which variable individually, P_1 or P_2 is a better predictor of F?
c. Which of the three models would you use to predict the final examination scores for a student who scored $P_1 = 78$ and $P_2 = 85$? What is your prediction in this case?

```r
library(broom)
dat = read.csv('../../data/exams.csv')

# a
mod1 = lm(F~P1, data=dat)
mod2 = lm(F~P2, data=dat)
```

```r
mod3 = lm(F~P1+P2, data=dat)

# b
# since both models have the same number of variables we can just use R-squared
summary(mod1)$r.squared
```

```
## [1] 0.8022501
```

```r
summary(mod2)$r.squared # winner
```

```
## [1] 0.8600357
```

```r
# c
# Now we need to compare across model sizes, adjusted R^2 or AIC (=Mallow's Cp) works here
mod1 %>% glance # Adj R2 .792, AIC 138
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.802         0.792  5.08      81.1 1.78e-8     1  -65.9  138.  141.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```r
mod2 %>% glance # Adj R2 .853, AIC 130
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.860         0.853  4.27      123. 5.44e-10     1  -62.1  130.  134.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```r
mod3 %>% glance # Adj R2 .874, AIC 128, winner
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.886         0.874  3.95      74.1 1.07e-9     2  -59.8  128.  132.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

RABE 3.14 + 4.7

A national insurance organization wanted to study the consumption of cigarettes in all 50 states and the District of Columbia. The data from 1970 are available on Canvas under `Files>data>cigarettes.csv`, and the variable definitions are given in the table below. For parts (a) and (b) below, specify the null and alternative hypotheses, the test used, and your conclusion using a significance $\alpha = .05$.

Variable	Definition
AGE	Median of the state's population
HS	Percentage of people over 25 years of age in a state who had completed high school
INCOME	Per capita personal income for a state (in dollars)
FEMALE	Percentage of population identified as "female"
PRICE	Average price (in cents) of a pack of cigarettes in the state

Variable	Definition
SALES	Number of packs of cigarettes sold in a state per capita

a. Test the hypothesis that the variable FEMALE is not needed in the regression equation relating SALES to the five predictor variables
b. Determine whether the variables FEMALE and HS should be included in the above regression equation
c. Compute that 95% Confidence Interval for the true regression coefficient of the variable INCOME.
d. What percentage of the variation in SALES can be accounted for by the three variables PRICE, AGE, and INCOME?
e. Using an added variable plot, show the effect of including the INCOME variable
f. What percentage of the variation in SALES can be accounted for when INCOME is removed from the above regression?
g. Compute the pairwise correlation coefficients matrix and construct the corresponding scatter plot matrix.
h. Are there any disagreements between the pairwise correlation coefficients and the corresponding scatter plot matrix?
i. Is there any difference between your expectations in part (a) and what you see in the pairwise correlation coefficients matrix, or in the scatter plot matrix?

```
dat = read.csv('../../data/cigarettes.csv')

# a
# H0: beta_FEMALE = 0, HA: beta_FEMALE =/= 0, alpha=.05
# we can test this hypothesis using a t-test on just the FEMALE coefficient
mod.full = lm(sales~., data=dat%>%select(-state))
mod.full %>% summary
```

```
##
## Call:
## lm(formula = sales ~ ., data = dat %>% select(-state))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.080 -11.396  -6.562   4.891 131.219
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  43.48020  230.47220   0.189  0.85121
## age           3.35027    2.78278   1.204  0.23491
## hs           -0.41080    0.65785  -0.624  0.53549
## income        0.02299    0.00856   2.686  0.01010 *
## female        0.98494    4.79326   0.205  0.83812
## price        -3.38367    1.01115  -3.346  0.00166 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.03 on 45 degrees of freedom
## Multiple R-squared:  0.3125, Adjusted R-squared:  0.2361
## F-statistic: 4.091 on 5 and 45 DF,  p-value: 0.003799
```

```
# t-stat is ~.21 and p-value is .838, fail to reject H0
```

```r
# b
# As with 3.3 compare across model sizes, adjusted R^2 works here
mod.reduced = lm(sales~ age+income+price, data=dat%>%select(-state))

summary(mod.full)$r.squared
```

```
## [1] 0.3125306
```

```r
summary(mod.reduced)$r.squared
```

```
## [1] 0.3032434
```

RABE 4.1a

Using the milk production dataset (on canvas under `Files>data>milk_production.csv`, described in RABE pages 3-4), fit the following model:

$$\text{CURRMILK} = \beta_0 + \beta_1\text{PREVIOUS} + \beta_2\text{FAT} + \beta_3\text{PROTEIN} + \beta_4\text{DAYS} + \beta_5\text{LACTAT} + \beta_6\text{I79}$$

Now, for your fit model determine:

- If the regression assumptions (linearity and iid normal errors) are met
- If any outliers are present in the data
- If any linear dependence exists between the independent variables