ProblemSet7

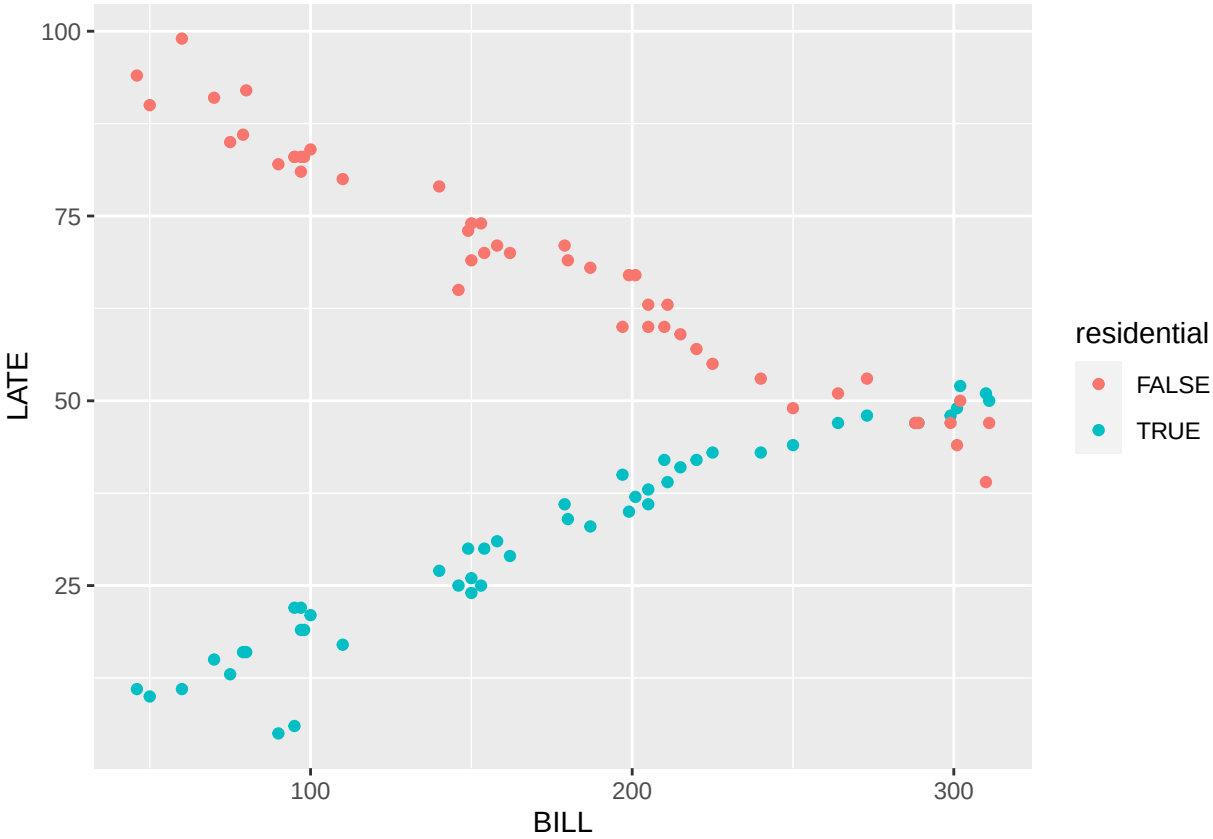
Aaron Graybill

4/5/2021

Problem 1.

I think it would be wise to start with a simple visualization of how the size of the loan affects the number of days to repayment. That plot is as follows:

```
## here() starts at /Users/aarongraybill/Documents/Haverford Stuff/Math/Math286
```



Well wow! That indicates that we definitely need to treat residential loans differentially. And simply having a different intercept will not suffice either, we will need to use an interaction term. This is nice data!

The table below summarizes the three possible regressions with and without interaction and dummy intercepts.

The output below makes it clear that the full model with the interaction term has the highest adjusted R^2, but we can further quantify whether or not the model is adding significant predictive power with a partial F test.

```
## Analysis of Variance Table
##
```

Table 1:

	Dependent variable:		
	LATE		
	(1)	(2)	(3)
BILL	−0.013	−0.013	−0.191***
	(0.031)	(0.019)	(0.006)
residential		−37.396***	−99.549***
		(2.944)	(1.695)
BILL:residential			0.357***
			(0.009)
Constant	51.984***	70.682***	101.758***
	(5.964)	(3.913)	(1.199)
Observations	96	96	96
R^2	0.002	0.635	0.980
Adjusted R^2	−0.009	0.627	0.980
Residual Std. Error	23.724 (df = 94)	14.421 (df = 93)	3.371 (df = 92)
F Statistic	0.163 (df = 1; 94)	80.910*** (df = 2; 93)	1,523.848*** (df = 3; 92)
Note:			*p<0.1; **p<0.05; ***p<0.01

```
## Model 1: LATE ~ BILL
## Model 2: LATE ~ BILL + residential
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     94 52904
## 2     93 19342  1     33563 161.38 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: LATE ~ BILL + residential
## Model 2: LATE ~ BILL * residential
##   Res.Df     RSS Df Sum of Sq    F    Pr(>F)
## 1     93 19341.7
## 2     92  1045.5  1     18296 1610 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

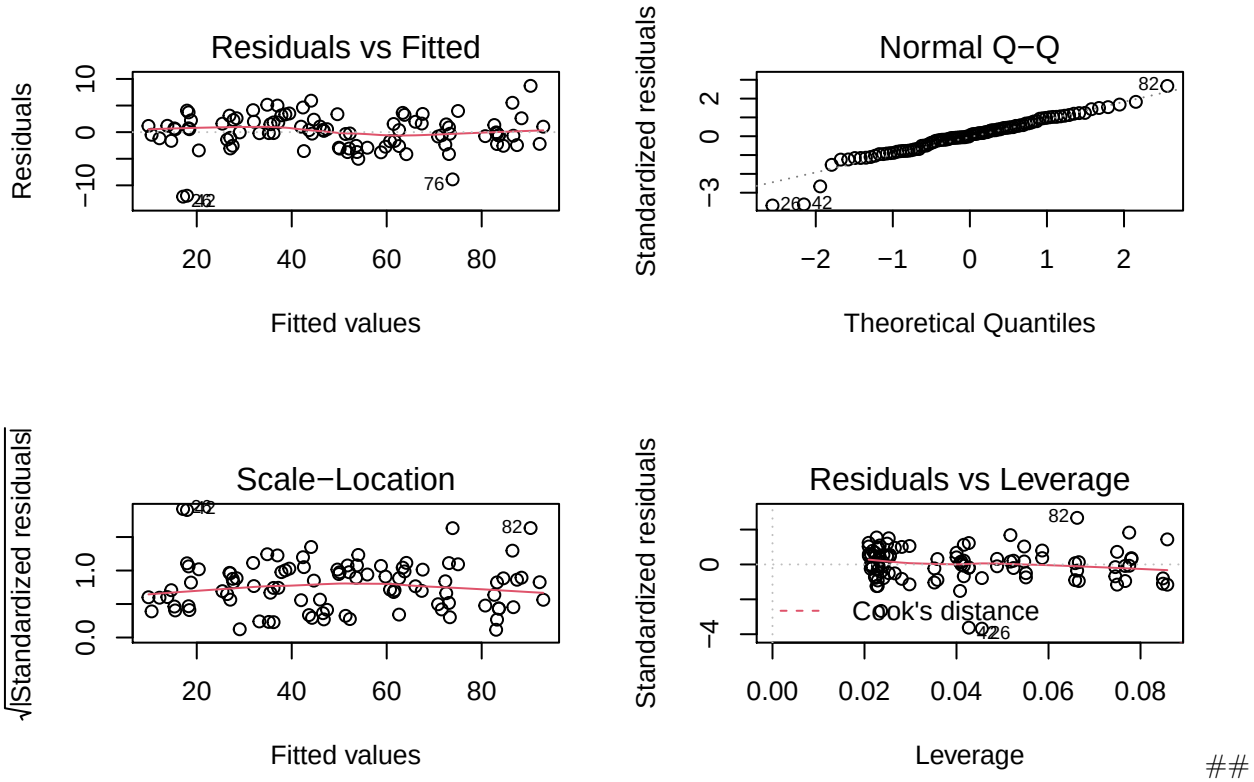
```
## Analysis of Variance Table
##
## Model 1: LATE ~ BILL
## Model 2: LATE ~ BILL * residential
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     94 52904
## 2     92  1045  2     51859 2281.7 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above output shows that in every case, going for the more complicated model adds substantial predictive power. Increasing from the model with no dummies to a dummy intercept is signifcant. Increasing from a model with dummies intercept to dummy interaction is significant. And of course, going from the simplest to the most complicated is also significant. As such we should select the model with the dummy interaction.

To summarize, we use the following model to predict the lateness of pay given bill size:

$$\texttt{LOAN} = \beta_0 + \beta_1 \texttt{BILL} + \beta_2 \texttt{residential} + \beta_3 \texttt{residential} \cdot \texttt{BILL}$$

To ensure that this is a suitable choice the diagnostic plots below show that none of the assumptions are flagrantly violated.



##

Problem 2.

a.

An ancova is inappropriate to decide whether or not an increase in low income percent is associated with an increase in percentage of grade one repeats. Instead a one way t-test of the coefficient on the low income students term.

```
##
## Call:
## lm(formula = X.Repeating.1st.Grade ~ X.Low.income.students, data = d)
##
## 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 ***
```

```
## X.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
```

The output above indicates that p-value on the one sided t-test would be 3.24×10^{-5} which says that an increase ini low income students students is significantly associated with an increase in the percent repeating 1st grade. This, of course, is when we control for no other factors (as the question asks).

b.

Again I don't really think an ancova is the right test for seeing if there has been a significant increase in the percentage of low income students from the 90s to the 2000s. Instead a two sample t-test on the sample means is the most appropriate. The results of such a test are as follows:

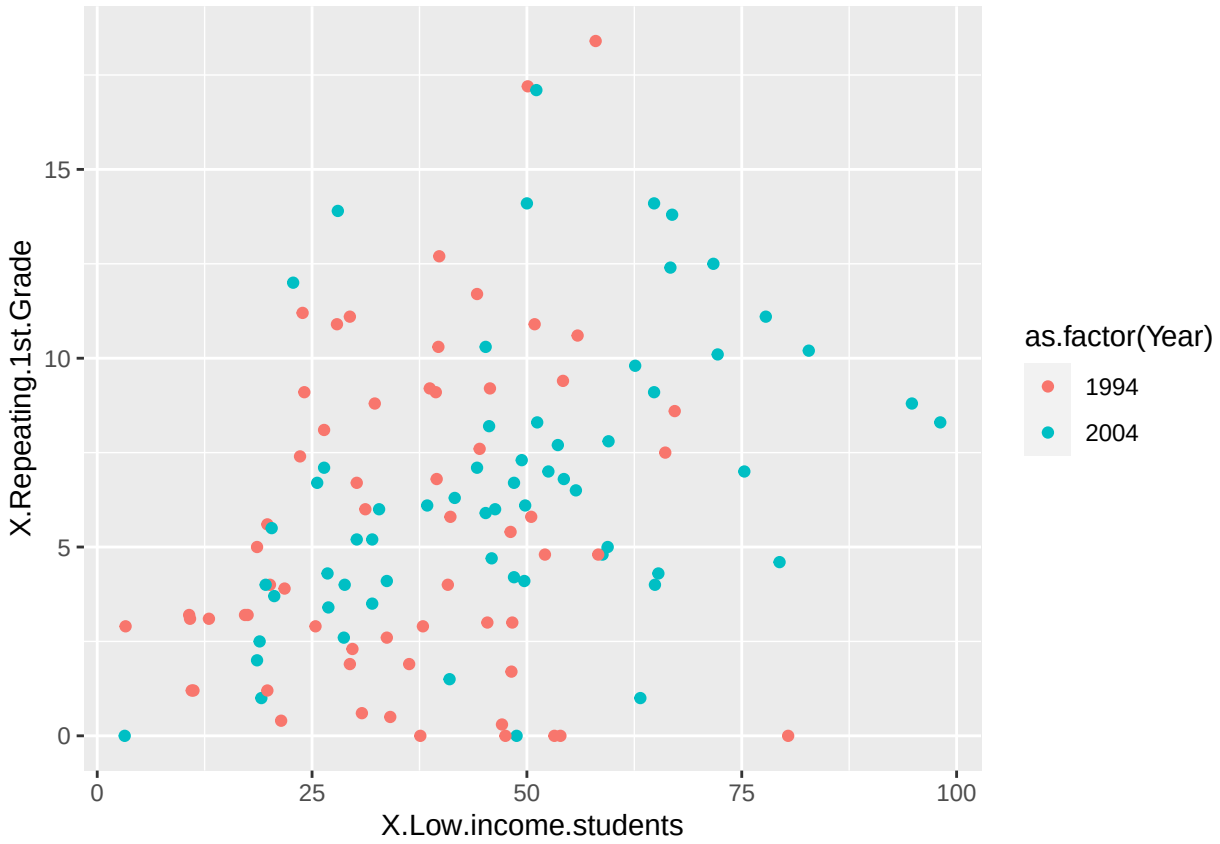
```
##
##  Welch Two Sample t-test
##
## data:  d$X.Low.income.students[d$Year == 2004] and d$X.Low.income.students[d$Year == 1994]
## t = 3.4301, df = 114.11, p-value = 0.0004203
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  5.85641     Inf
## sample estimates:
## mean of x mean of y
##  47.54918  36.21148
```

The test above shows that there is strong evidence ($p = 0.0004203$) that the mean percentage of students in poverty increased between 1994 and 2004 (which is pretty devastating).

c.

When it says any association that gives us the freedom to use a fully dummy-with-interaction model to perform the ancova, we will test the simplest baseline model with no year dummy against the model with the intercept and interaction from the year dummy. That ancova output is:

```
## Analysis of Variance Table
##
## Model 1: X.Repeating.1st.Grade ~ X.Low.income.students
## Model 2: X.Repeating.1st.Grade ~ X.Low.income.students * as.factor(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
```



The ancova output above and the corresponding plot indicate that we have insufficient evidence to say that the percentage of low income students has different impact on the percentage of students repeating grade one in 1994 vs 2004. Basically, the relationship between the two seems to be similar across time. The plot seems to echo this fact, although the datasets are in different locations, their slope and intercepts seem to remain the mostly the same.

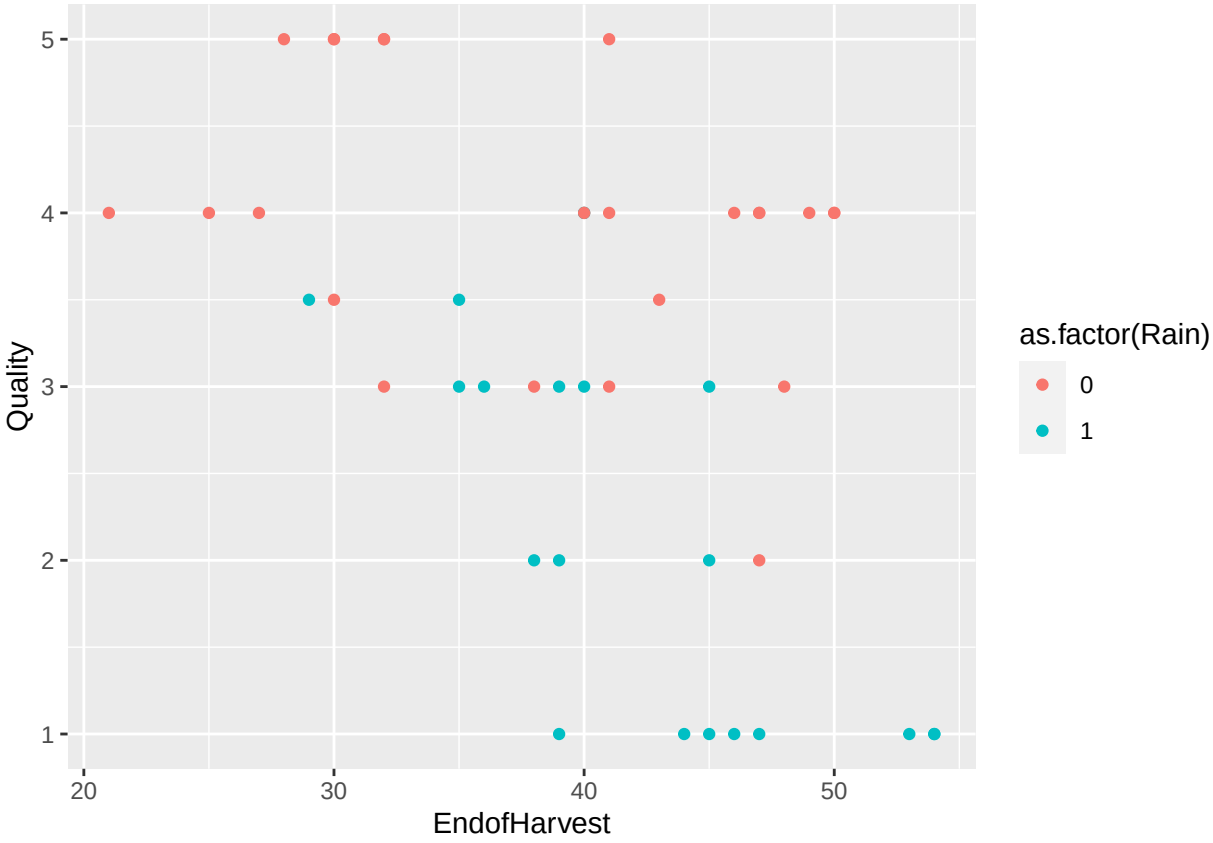
The headline that the reading standard is too hard seems to be in affected by poverty in mostly the same way to previous years, so the reading standard is likely not to blame.

Problem 3.

```
##
## Call:
## lm(formula = Quality ~ EndofHarvest * Rain, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6833 -0.5703  0.1265  0.4385  1.6354
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.16122    0.68917   7.489 3.95e-09 ***
## EndofHarvest       -0.03145    0.01760  -1.787   0.0816 .
## Rain                1.78670    1.31740   1.356   0.1826
## EndofHarvest:Rain -0.08314    0.03160  -2.631   0.0120 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.7578 on 40 degrees of freedom
## Multiple R-squared:  0.6848, Adjusted R-squared:  0.6612
## F-statistic: 28.97 on 3 and 40 DF,  p-value: 4.017e-10
```

The regression output given above shows that the last term, the interaction term, is signifcant at the 5% which implies that unwanted rain at harvest is not a single shock to the quality of the wine and instead affects wines that would have been worse quality more negatively than those that would have been good quality wines. Specifically, if rain happens and the number of days since August 31st is large (bad wines), the interaction term begins to have a large negative effect on the quality. Whereas for quickly harvested wines, the rain has less of an impact on the quality. One could summarize this with, late rains make worse wines than early rains, holding other factors constant.



b.

i. The equation for no rain at harvest is:

$$\hat{Q} = 5.16122 - 0.03145t$$

Where $\hat{Q}$ is the estimated quality and t is number of days. Therefore, the derivative of $\hat{Q}$ wrt t is -0.03145. To compute a one unit decrease in quality, we simply take $\frac{1}{0.03145}$ giving 31.8 days after August 31st. And by linearity, every 31.8 days after that will decrease quality by another point.

ii. The equation with rain is:

$$\hat{Q} = (5.16122 + 1.78670) + (-0.03145 - 0.08314)t = 6.94792 - 0.11459t$$

Computing a one unit decrease in $\hat{Q}$ can be done another way, as follows:

$$\begin{aligned} -1 &= \hat{Q}_t - \hat{Q}_0 \\ &= 6.94792 - 0.11459t - (6.94792 - 0.11459 \cdot 0) \\ &= -0.11459t \end{aligned}$$

Solving this gives that $t = 8.726765$, so the quality loses one star every 8.73 days! This is a much faster quality drop off than no-rain grapes.