Regression Analysis
Final Exam

Name:_____

Take the allotted time to complete this exam. You may use writing utensils, calculator, and computer for using R only. Absolutely no phones, apps, notes, books, etc. allowed. You are on your honor to do your own work. Good luck!

1. [**40 POINTS**] ACME University administrators want to predict the total number of credit hours a student will complete based on their majors in college, their ACT (standardized admission test) scores, their high school GPAs, and their heights in inches. Data on twelve randomly sampled students from last year are below.

StudentID	Major	ACT	Height.inches	Credit.Hours
1	Math	32	66	121
2	Math	35	69	144
3	Math	28	70	137
4	Math	26	71	152
5	Math	29	64	122
6	Math	34	74	145
7	Chem	27	65	120
8	Chem	26	63	128
9	Chem	29	71	120
10	Chem	25	68	122
11	Chem	24	67	132
12	Chem	31	67	125

   (a) [**5 POINTS**] Write down the model matrix for these data (the X matrix) for finding the least-squares coefficient estimates for a linear model for predicting the total number of credit hours a student will take based on major (Math or Chem), ACT score, and height in inches.

   (b) [**5 POINTS**] Write down the Y vector for these data used to find the least-squares coefficient estimates.

(c) **[5 POINTS]** The linear model output from R is below:

```
> out <- lm(Credit.Hours ~ Major + ACT + Height.inches, data)
> summary(out)

Call:
lm(formula = Credit.Hours ~ Major + ACT + Height.inches, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5287  -6.3170   0.2087   5.7811  10.3511

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    24.0662    59.4368   0.405   0.6962
MajorMath      11.1286     6.1092   1.822   0.1060
ACT            -0.7349     0.8985  -0.818   0.4371
Height.inches   1.7996     0.8827   2.039   0.0758 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.7 on 8 degrees of freedom
Multiple R-squared:  0.5646,    Adjusted R-squared:  0.4014
F-statistic: 3.458 on 3 and 8 DF,  p-value: 0.07128
```

Based on this output, **AND** based on common sense, which variable(s) do you suggest we keep in the model and which could possibly be discarded? Explain. Remember our goal is to be able to predict the number of credit hours a student will take in college.

(d) **[5 POINTS]** The pair-wise correlations for all three variables, *including the response*, are

```
> cor(data[,3:5])
```

```
                    ACT Height.inches Credit.Hours
ACT           1.0000000     0.3405461    0.2247371
Height.inches 0.3405461     1.0000000    0.6195331
Credit.Hours  0.2247371     0.6195331    1.0000000
```

Based on this, are there signs of collinearity in the predictor variables? Explain.

(e) **[5 POINTS]** Use the full model output from R in part (c) above to predict the total number of credit hours a student will take in college who is 70 inches tall, had an ACT score of 30, and who majors in math.

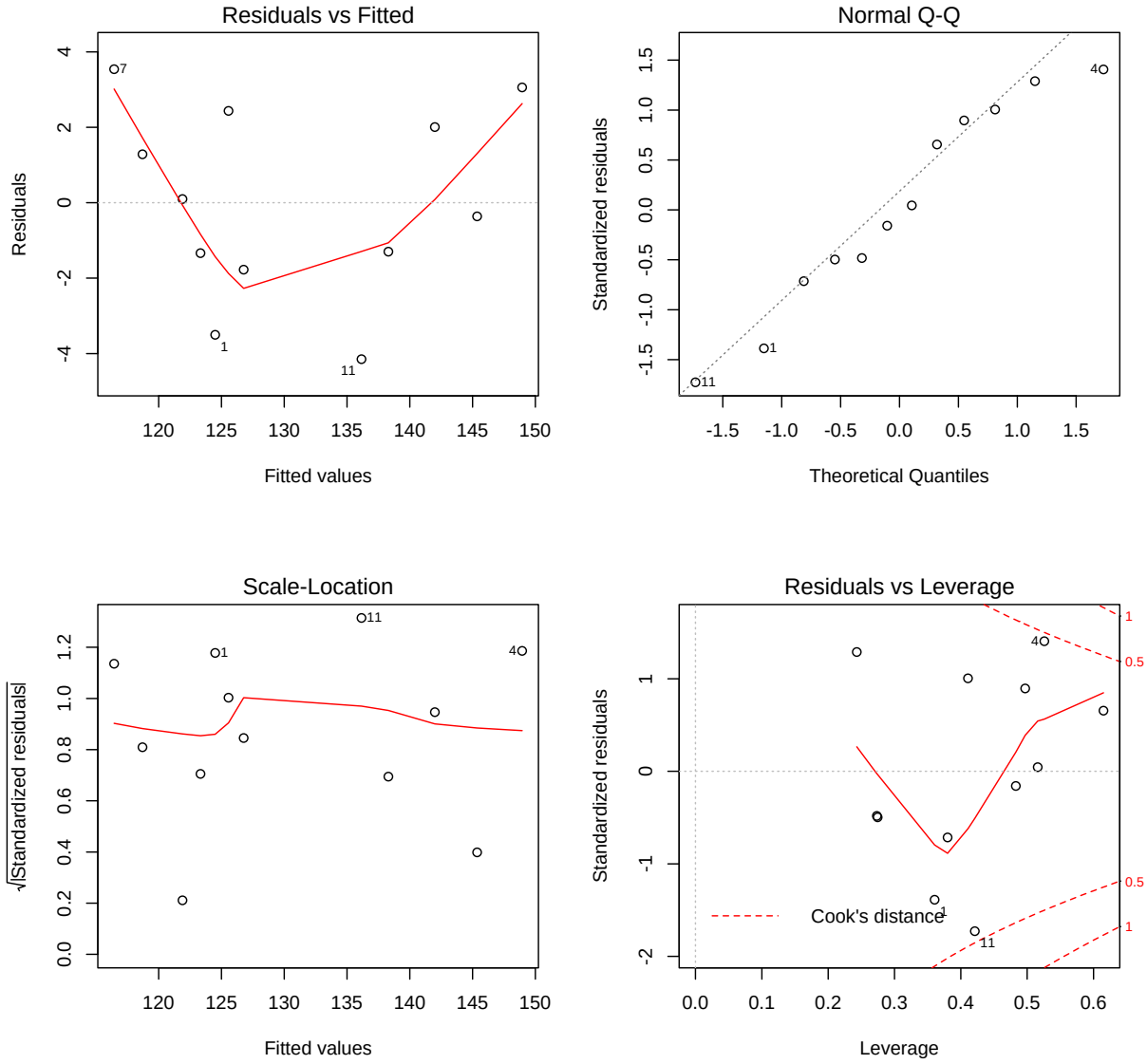
(f) **[5 POINTS]** Use the full model output from R in part (c) above to predict the total number of credit hours a student will take in college who is 66 inches tall, had an ACT score of 28, and who majors in chemistry.

(g) **[5 POINTS]** A new variable, the number of extra curricular activities in high school (measured in years) is discovered, and the output for a new model is given below:

```
> out2 <- lm(Credit.Hours ~ Major + ACT + Height.inches + Extra.Cur.Yrs, data)
> summary(out2)

Call:
lm(formula = Credit.Hours ~ Major + ACT + Height.inches + Extra.Cur.Yrs,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1485 -1.4460 -0.1312  2.1128  3.5405

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   105.22989   24.23350   4.342 0.003386 **
MajorMath      -0.68923    2.73980  -0.252 0.808605
ACT             0.25898    0.35295   0.734 0.486937
Height.inches  -0.02086    0.40513  -0.051 0.960374
Extra.Cur.Yrs   1.86433    0.25417   7.335 0.000158 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.156 on 7 degrees of freedom
Multiple R-squared:  0.9499,    Adjusted R-squared:  0.9212
F-statistic: 33.16 on 4 and 7 DF,  p-value: 0.0001219
```

Remark on which variables now seem important, what has changed and why. Specifically address (1) the change in p-values for the t-tests, (2) the new Multiple/Adjusted R-squared values, and (3) the values of the beta estimates. Also (4) which (if any) variables would you suggest dropping from the model?

(h) **[5 POINTS]** Use the diagnostic plots below to remark on whether or not the assumptions have been adequately met. If anything seems wrong, be sure to suggest some kind of appropriate remedial measure.



~~~~~~~~~~ *End of the Credit Hours Problem Questions* ~~~~~~~~~~

2. [**5 POINTS**] Write down the *general matrix equation* for finding the least-squares coefficient estimates. Don't write any data values here: write the matrix formula.

3. [**5 POINTS**] Write down the general matrix equation for the hat matrix H in terms of the X matrix.

4. [**5 POINTS**] Write down matrix equation for the residuals in terms of the hat matrix H and the matrix Y.

5. [**5 POINTS**] Write down the formula for the linear correlation coefficient r when there is just one predictor variable.

6. [**5 POINTS**] Write down the simple little formula (in terms of r and sample standard deviations) for the simple linear regression slope estimate when there is just one predictor variable.

7. [**5 POINTS**] What are the assumptions that need to be checked when constructing a linear model?
~~~~~~~~~~

8. **[15 POINTS]** Complete the following lack-of-fit ANOVA table:

Source	DF	SS	MS	F^*	p-value
Regression		34.783			
Residual				NA	NA
Lack-of-Fit	5				
Pure Error		2.110		NA	NA
Total	21	41.85	NA	NA	NA

9. **[5 POINTS]** Which is used in the case of nonconstant error variance when the errors are uncorrelated?

   (a) Durbin-Watson test (b) Shapiro-Wilk test (c) weighted least-squares (d) none of these

10. **[5 POINTS]** Which is to be used when errors are found to be correlated, and can also help mitigate nonconstant error variance?

    (a) Durbin-Watson (b) Shapiro-Wilk (c) generalized least-squares (d) none of these

11. **[5 POINTS]** How can you check for correlated residuals?

    (a) Durbin-Watson (b) run cor() on 'neighboring' residuals
    (c) plot residuals vs. fits (d) all of these

12. **[5 POINTS]** How can you check for normality of residuals?

    (a) Anderson-Darling (ad.test) (b) Shapiro-Wilk (shapiro.test)
    (c) normal probability plot (qqnorm) (d) all of these

13. **[5 POINTS]** How can you check for equal variance of the residuals?

    (a) plot residuals vs. fits;

    (b) Levene's test on a partitioning of the data;

    (c) regress $\sqrt{|residuals|}$ vs. predictor variable and check p-value associated with the slope;

    (d) all of these.

14. **[5 POINTS]** True or False: Cook distances measure the changes in fitted values when removing data points and so are used to identify potential influential observations.

15. [**20 POINTS**] The four problems on the next page refer to the following R output.

```
> model <- lm(Employed ~ ., data = longley)
> summary(model)

Call:
lm(formula = Employed ~ ., data = longley)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41011 -0.15767 -0.02816  0.10155  0.45539

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.482e+03  8.904e+02  -3.911 0.003560 **
GNP.deflator  1.506e-02  8.492e-02   0.177 0.863141
GNP          -3.582e-02  3.349e-02  -1.070 0.312681
Unemployed   -2.020e-02  4.884e-03  -4.136 0.002535 **
Armed.Forces -1.033e-02  2.143e-03  -4.822 0.000944 ***
Population   -5.110e-02  2.261e-01  -0.226 0.826212
Year          1.829e+00  4.555e-01   4.016 0.003037 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3049 on 9 degrees of freedom
Multiple R-squared:  0.9955,    Adjusted R-squared:  0.9925
F-statistic: 330.3 on 6 and 9 DF,  p-value: 4.984e-10


> x <- model.matrix(model)[,-1]
> e <- eigen(t(x)%*%x)
> e$val

[1] 6.665299e+07 2.090730e+05 1.053550e+05 1.803976e+04
[5] 2.455730e+01 2.015117e+00


> sqrt(e$val[1]/e$val)

[1]    1.00000   17.85504   25.15256   60.78472 1647.47771
[6] 5751.21560


> vif(x)

NP.deflator          GNP   Unemployed Armed.Forces
   135.53244   1788.51348     33.61889      3.58893
  Population         Year
   399.15102    758.98060
```

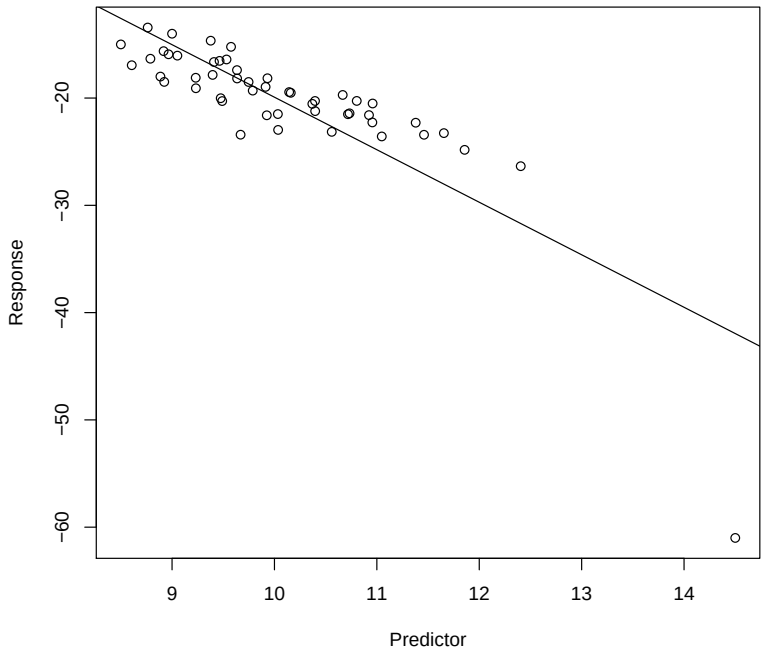
(a) **[5 POINTS]** In the output on the previous page, circle the condition numbers.

(b) **[5 POINTS]** Do the condition numbers seem to indicate anything about some of the variables in particular? Explain.

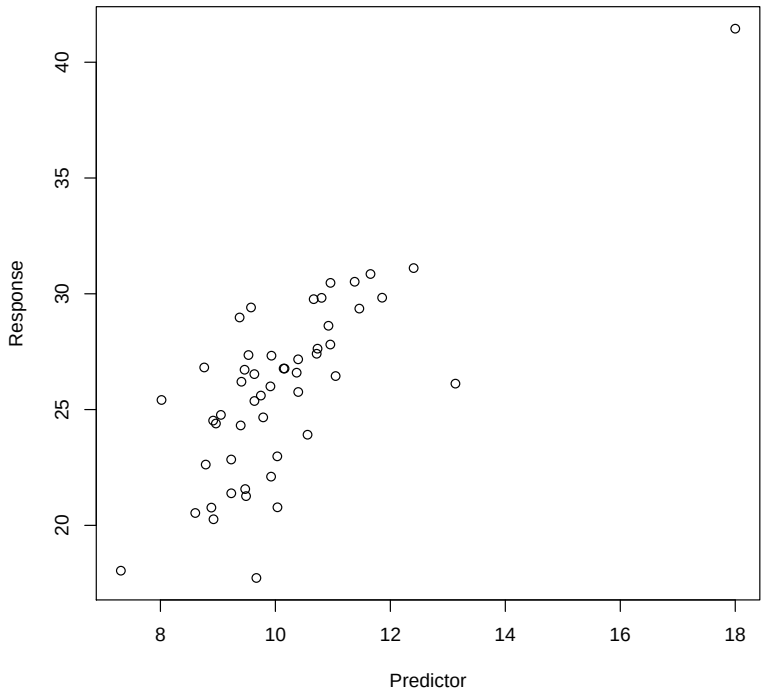
(c) **[5 POINTS]** In the output on the previous page, draw a rectangle around the variance inflation factors.

(d) **[5 POINTS]** Do the variance inflation factors seem to indicate anything about some of the variables in particular? Explain.

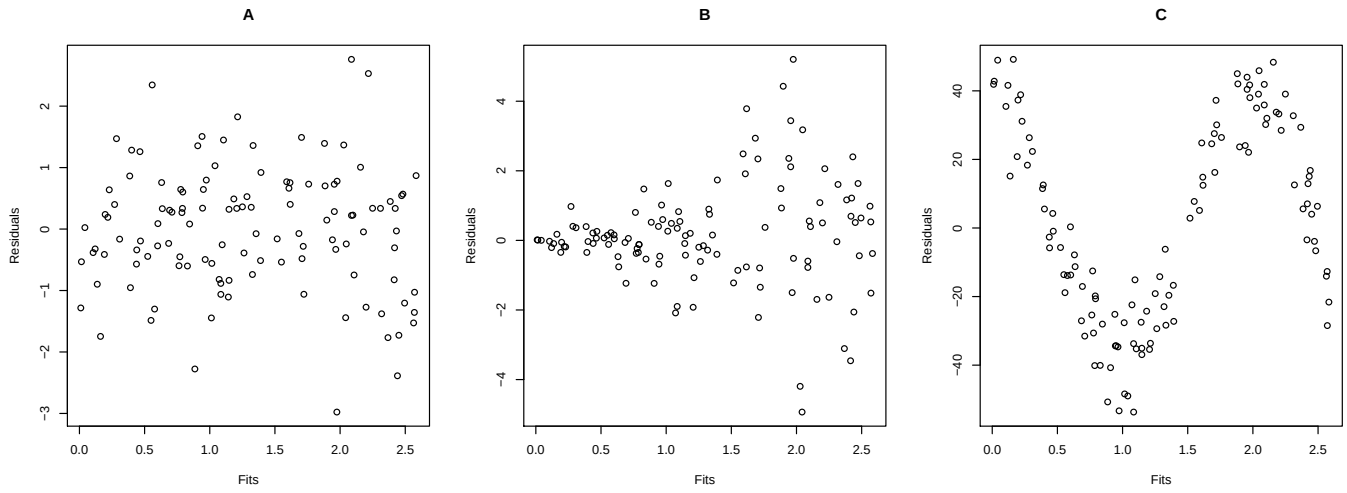
16. [**5 POINTS**] Circle the best candidate for an influential observation.



17. [**5 POINTS**] Circle the best candidate for a leverage point.



18. **[20 POINTS]** The following three plots were made in conjunction with three different linear models of the form $lm(y \sim x1 + x2)$.



Use the plots to answer these questions.

(a) **[5 POINTS]** Which plot (A, B, or C) most indicates constant error variance?
(a) A (b) B (c) C

(b) **[5 POINTS]** Which plot (A, B, or C) most indicates the error variance is increasing in the fitted values?
(a) A (b) B (c) C

(c) **[5 POINTS]** Which plot (A, B, or C) most indicates a nonlinear model might be more appropriate than the one made with $lm(y \sim x1 + x2)$?
(a) A (b) B (c) C

(d) **[5 POINTS]** How would you recommend correcting for the problem in plot B?

19. **[5 POINTS]** When doing principle components analysis (PCA) or principle components regression (PCR), why is it typically important to standardize the predictor variables?

20. **[5 POINTS]** What is the objective function to be minimized in ridge regression?

21. **[5 POINTS]** What is the objective function that is to be minimized in lasso regression?

22. **[5 POINTS]** How are the tuning parameters found when doing ridge and lasso regression?

23. **[5 POINTS]** Match the following link functions to their names:

$\eta(p) = \Phi^{-1}(p)$	Logit
$\eta(p) = \log(-\log(1-p))$	Probit
$\eta(p) = \log(p/1-p)$	Complementary log-log

24. **[5 POINTS]** Match the following goodness-of-fit measures to their names.

$1 - \frac{n-1}{n-p-1}\frac{SSE_p}{SSTO}$	Mallow'sC_p
$(y - Xb)^T(y - Xb)$	AIC
$SSE_p/MSE + 2p - n$	BIC (also called SBC)
$n\log(SSE_p) - n\log n + p\log n$	R^2_{adj}
$n\log(SSE_p) - n\log n + 2p$	SSE

25. **[5 POINTS]** What is the null deviance and what should you do if it is large?

26. **[5 POINTS]** What is the residual deviance and what should you do if it is large?

27. **[5 POINTS]** If the response is truly a binomial random variable, and the n_i are relatively large, what is the approximate distribution of the deviance?

28. [**20 POINTS**] Data and R output on the incidence of respiratory disease in infants by sex and feeding method are below.

	Bottle Only	Some Breast with Supplement	Breast Only
Boys	77/458	19/147	47/494
Girls	48/384	16/127	31/464

```
Call:
glm(formula = cbind(disease, nondisease) ~ sex + food, family = binomial,
    data = babyfood)

Deviance Residuals:
      1        2        3        4        5        6
 0.1096  -0.5052   0.1922  -0.1342   0.5896  -0.2284

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6127     0.1124 -14.347  < 2e-16 ***
sexGirl      -0.3126     0.1410  -2.216   0.0267 *
foodBreast   -0.6693     0.1530  -4.374 1.22e-05 ***
foodSuppl    -0.1725     0.2056  -0.839   0.4013
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26.37529  on 5  degrees of freedom
Residual deviance:  0.72192  on 2  degrees of freedom
AIC: 40.24

Number of Fisher Scoring iterations: 4
```

(a) [**5 POINTS**] Use the output to predict the probability that a baby girl who only bottle feeds will contract a respiratory disease.

(b) [**5 POINTS**] Use the output to predict the probability that a baby boy who only breast feeds will contract a respiratory disease.

(c) [**5 POINTS**] Breast feeding reduces the odds of respiratory disease to _____ % of that for bottle feeding.

(d) [**5 POINTS**] Is the χ^2 approximation valid for obtaining a p-value for the residual deviance in this case? Explain.

29. **[20 POINTS]** Santa Claus wants to estimate how many ounces of milk will be left at houses he visits based on (1) the assessed property value of the residence; (2) if the children of the household were naughty, nice, or some of each (mixed). Below are some data he took last year.

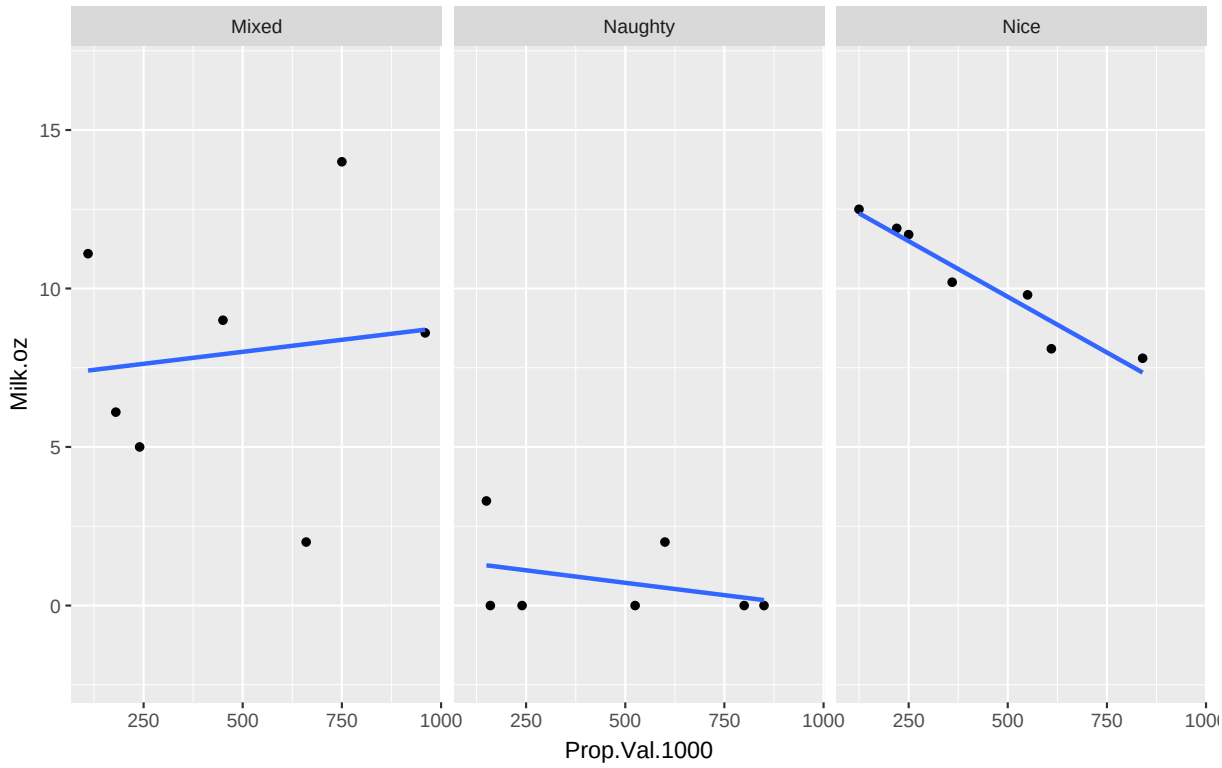
Milk (ounces)	Property Value ($1000)	Behavior
0.0	240	Naughty
0.0	160	Naughty
3.3	150	Naughty
0.0	525	Naughty
2.0	600	Naughty
0.0	800	Naughty
0.0	850	Naughty
11.9	220	Nice
10.2	360	Nice
12.5	125	Nice
8.1	610	Nice
7.8	840	Nice
9.8	550	Nice
11.7	250	Nice
11.1	110	Mixed
5.0	240	Mixed
9.0	450	Mixed
8.6	960	Mixed
6.1	180	Mixed
14.0	750	Mixed
2.0	660	Mixed

The R output for the full model with interactions is

```
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  7.2465705  1.8944850  3.8251 0.001656
Prop.Val.1000                0.0015146  0.0033608  0.4507 0.658670
BehaviorNaughty             -5.7444806  2.7584885 -2.0825 0.054832
BehaviorNice                 6.0066658  2.7938943  2.1499 0.048282
Prop.Val.1000:BehaviorNaughty -0.0030829  0.0049670 -0.6207 0.544125
Prop.Val.1000:BehaviorNice    -0.0085443  0.0054161 -1.5776 0.135515

n = 21, p = 6, Residual SE = 2.64862, R-Squared = 0.78
```

(a) **[5 POINTS]** According to the R output on the previous page, do the interaction terms seem statistically significant? Explain.

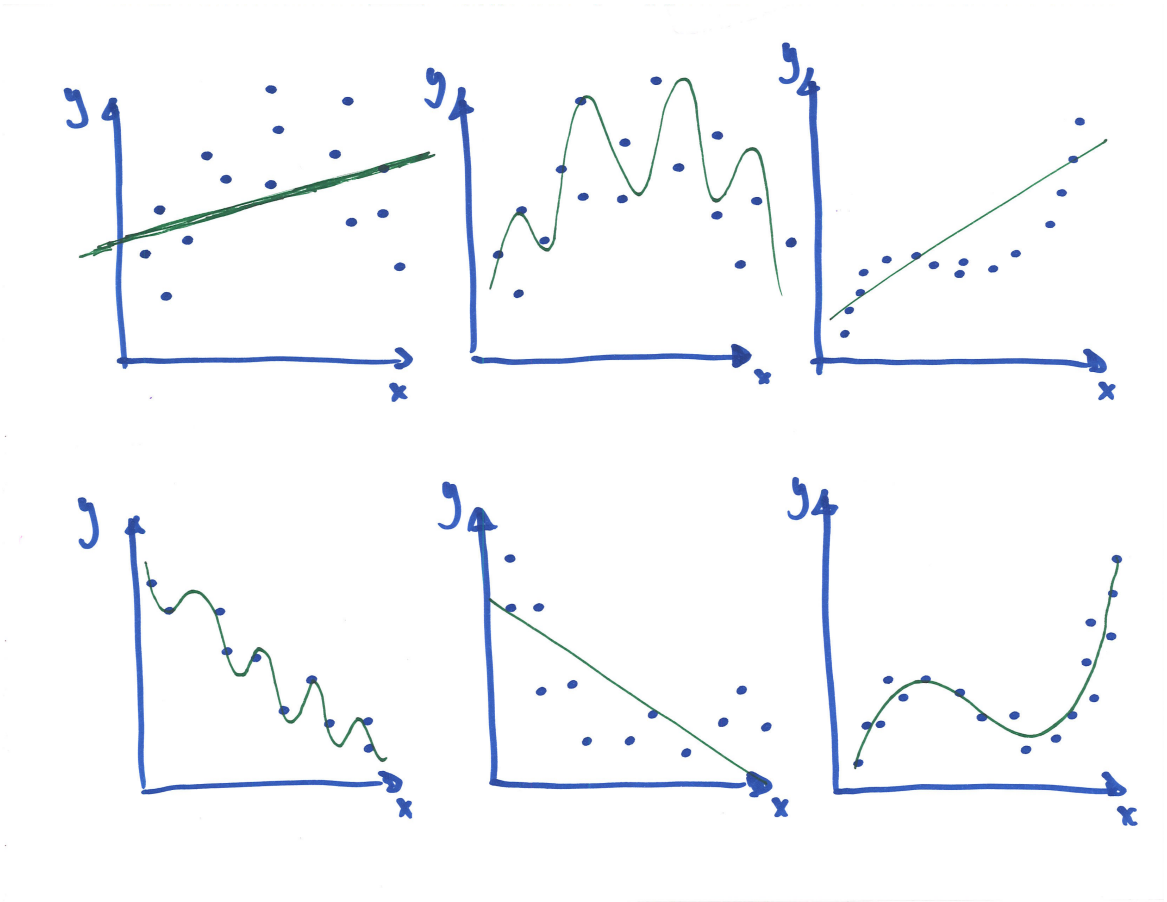
(b) **[5 POINTS]** Now considering the following plot of the data, do you think the interaction terms are significant? Explain.



(c) **[5 POINTS]** Use the model to predict the amount of milk left for Santa at a \$500,000 home of a child who was partly naughty, and partly nice (mixed).

(d) **[5 POINTS]** Use the model to predict the amount of milk left for Santa at a \$150,000 home of a nice child.

30. **[10 POINTS]** Which of the following modeling scenarios look (i) like a high bias scenario; (ii) like a high variance scenario; (iii) just about right? *Label each of the six plots with (i) or (ii) or (iii).*



31. **[5 POINTS]** What happens to the training error (such as SSE or MSE evaluated on the training set) as the model flexibility increases?

32. **[5 POINTS]** What happens to the test or cross validation error (such as SSE or MSE evaluated on the test or cross validation set) as the model flexibility increases?