MSiA 401: Predictive Analytics I

1. I was reviewing an article. The main hypothesis was whether hospitals with a religious affiliation (x) have lower perceived quality (y) than those with no religious affiliation. The authors gathered a sample of over 500 health systems in the US and carefully classified them into those with or without religious affiliation. They merged this with a large consumer survey evaluating quality of the systems. You may assume that the survey methodology was valid. They found a significant difference in the mean perceived quality level between the two groups, with non-religious hospitals having higher perceived quality. What can we conclude about the effect of religious affiliation on perceived quality? Explain. *Answer: Beware of the omitted variable. There could be a w that causes both x and y. The best answers would give a plausible example w, such as having a university affiliation—perhaps having a university affiliation is the cause of perceived quality, and systems affiliated with universities are less likely to have a religious affiliation (although those with religious and university affiliations have high perceived quality). Or perhaps it is due to advertising and public relations (PR) focused on quality. Those that do ad/PR have higher perceived quality, while those with religious affiliations do less ad/PR or brand themselves differently.*

2. I am interested in mapping a certain atmospheric property (call it y in this problem) to temperature (degrees Celsius, labeled `degC`), i.e., for a given temperature I want to know y. I obtained data from a table in the engineering toolbox, but no formula was provided. I generated the graph shown on the next page (see Raw data: y vs. `degC` in the upper left) and observed a relationship that looked roughly exponential. I therefore decided that a log transform of y might be appropriate, and plotted $\log(y)$ against `degC` (upper right). Output from the estimated regression is provided below.

```
> summary(fit)
Call: lm(formula = log(y) ~ degC, data = dat2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.377350   0.029786   46.24   <2e-16 ***
degC        0.074356   0.001348   55.16   <2e-16 ***

Residual standard error: 0.1472 on 23 degrees of freedom
Multiple R-squared:  0.9925,Adjusted R-squared:  0.9922
F-statistic:  3043 on 1 and 23 DF,  p-value: < 2.2e-16
```

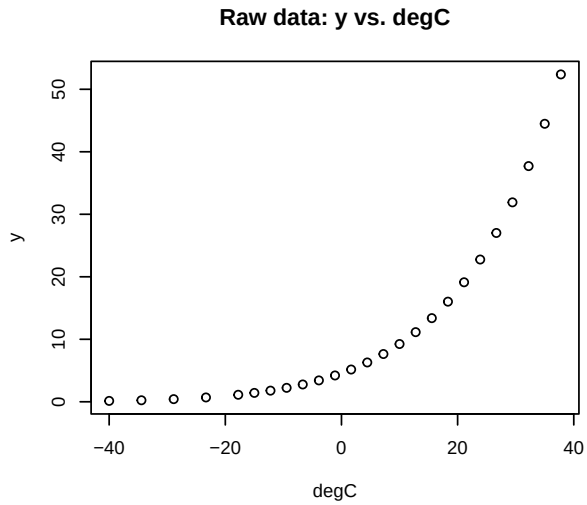
   (a) What does $R^2 = 0.9925$ tell you? *Answer: Knowing the value of temperature explains 99.25% of the variation in* $\log(y)$*, which suggests a good fit at least within this range of temperatures.*

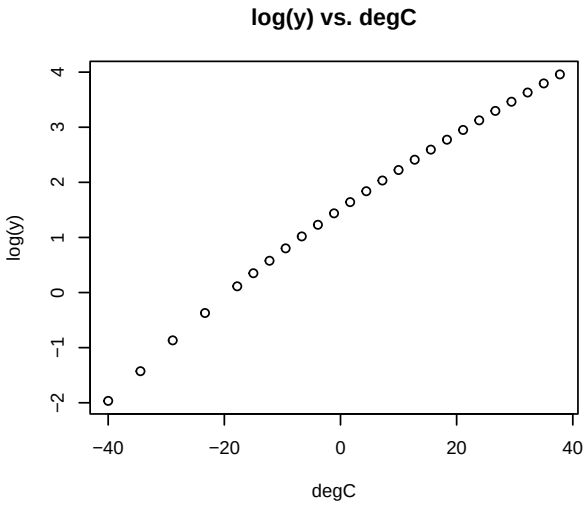
(b) Examine the diagnostic plots and discuss the quality of the fit, referencing specific plot(s). *Answer: The residual plot shows a clear pattern indicating that this model does not fit the data, even though it explains 99.25% of the variation in this range of temperatures.*

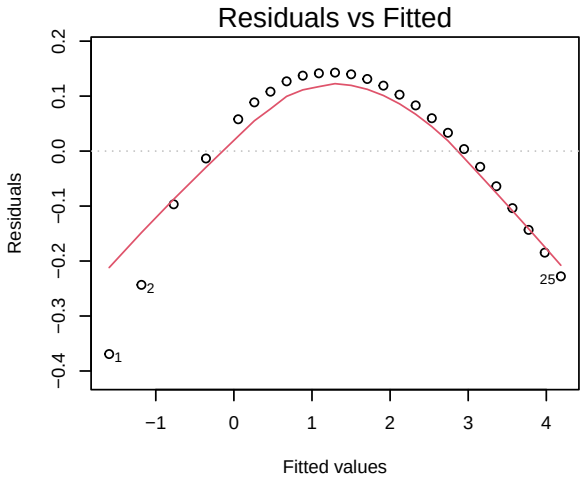
(c) I wish to know y when the temperature is 21°C. Would you recommend using this regression model? Explain briefly. For my purpose the answer does not need to be exact, but should be "in the ballpark." *Answer: The near-perfect value of R^2 suggests that the predictions will be very good within this range of temperatures (interpolation). In fact, I tried a non-parametric method to get the right value (2.9441 versus 2.9388, so I'm off by less than 0.2% in log values. After exponentiating, 18.89366 versus 18.99371, or about 0.5%. The log model is accurate to three significant digits.).*

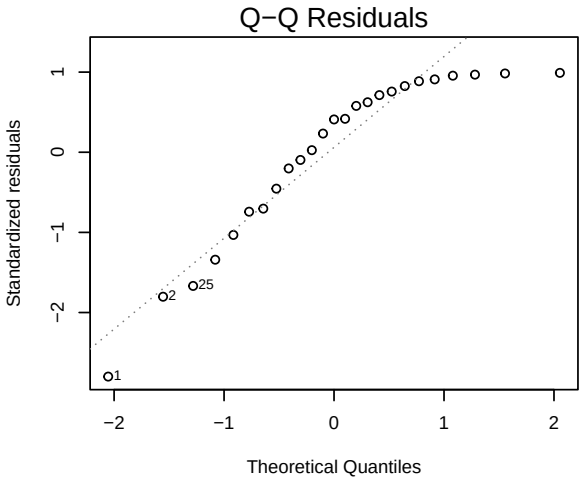
(d) If I wanted a 95% prediction interval for my estimate in part 2c, discuss whether I could use the predict function in R, which is $\hat{y} \pm 2.07 S_{Y|x_0=21}$, where $\hat{y}$ is the predicted value, 2.07 is the 0.975 percentile of a t distribution with 23df, and $S_{Y|x_0=21}$ is the standard error of a new observation Y given $x_0 = 21$. *Answer: There are two ways to answer this question. The PI formula assumes a normal distribution, and the CLT does not help for PIs. The residual plot indicates non-normal residuals, especially with problems in the tails. Students should indicate that the assumptions are violated, although I have not been specific about what is meant by "in the ballpark" so give full credit for either yes or no. The important thing is to note that (a) the residuals are not normal and (2) the PI formula requires normality. Alternatively for full credit, the model is not right so don't trust the PI (the standard error is estimating the lack of fit rather than the irreducible error). The PI is, in fact, (13.80, 25.88), which actually includes the true value, 18.99.*

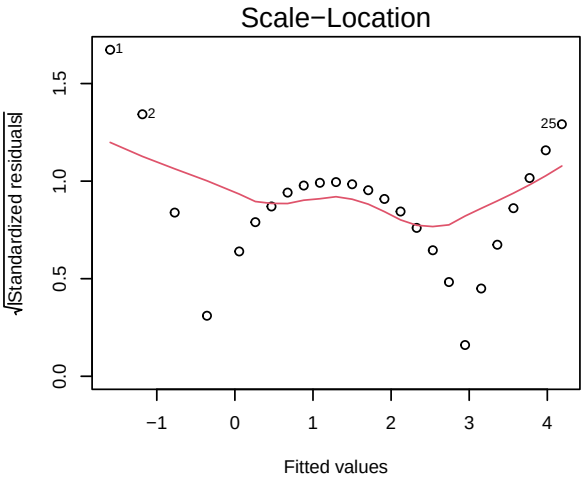
(e) Atmospheric temperatures decrease with elevation. Would you recommend using this regression model to know y for a temperature of −80°C? Explain briefly *Answer: We are extrapolating well outside the range of the observed data with a function that does not fit. I would not use this regression model.*

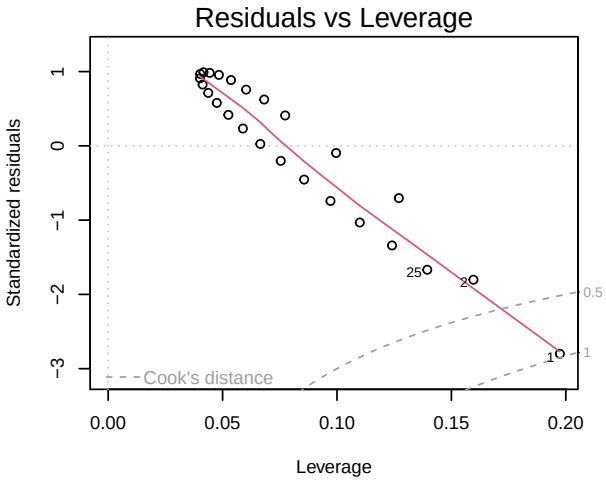

Raw data: y vs. degC
y
degC


log(y) vs. degC
log(y)
degC


Residuals vs Fitted
Residuals
Fitted values


Q−Q Residuals
Standardized residuals
Theoretical Quantiles


Scale−Location
√|Standardized residuals|
Fitted values


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance

3. (3 points) Next quarter you will study methods that automatically learn the regression function without you having to specify transformation terms prior to estimation. An example is shown to the right (solid black line/curve). How do you suggest quantifying the uncertainty about the regression line/curve (dotted lines/curves)? Circle the **best** answer.

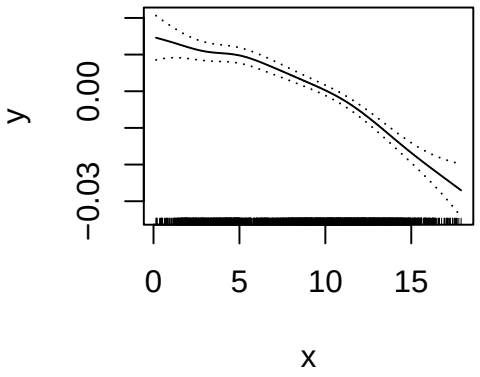
   (a) Confidence interval (CI) for $\hat{\beta}_1$

   (b) CI for the mean prediction

   (c) Prediction interval

   (d) CI for S_e^2

Answer: CI for the mean prediction



4. (10 points) The salary of workers (in thousands of dollars) is regressed on the age of the worker, the number of years of experience, gender (female=1 for females and 0 for males), training level (low, medium or high), and the interaction between gender and training levels. The output is below. I added a column `j` that gives the subscript of β_j to be used in the statement of hypotheses below, e.g., β_0 is the intercept parameter, β_1 is the slope parameter for `ageyrs`, etc. You may assume that the residual, QQ and Cook's distance diagnostic plots look

```
> fit = lm(salary/1000 ~ ageyrs + expyrs + female * train, employee)
> drop1(fit, test="F)
             Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>                    2337.7 264.09
ageyrs        1    316.05 2653.7 271.09  ??????  0.004872 **
expyrs        1    693.04 3030.7 280.52  ?????? 5.607e-05 ***
female:train  2    396.54 2734.2 271.21  ??????  0.007186 **

> summary(fit)

Call: lm(salary/1000 ~ ageyrs + expyrs + female * train, employee)
j                  Estimate Std. Error t value Pr(>|t|)
0 (Intercept)       24.2158     4.9426   4.899 7.04e-06 ***
1 ageyrs             0.3221     0.1104   2.918  0.00487 **
2 expyrs             1.0848     0.2510   4.322 5.61e-05 ***
3 female           -13.8068     2.2489  -6.139 6.14e-08 ***
4 trainMed           5.4780     2.1952   2.495  0.01521 *
5 trainHigh         10.1405     2.9978   3.383  0.00124 **
```

```
6 female:trainMed    8.9347     3.3327   2.681  0.00936 **
7 female:trainHigh  12.1459     4.6795   2.596  0.01174 *

Residual standard error: 6.091 on 63 degrees of freedom
Multiple R-squared:  0.7146,Adjusted R-squared:  0.6829
F-statistic: 22.53 on 7 and 63 DF,  p-value: 6.01e-15
```

(a) (3 points) What null and alternative hypotheses are tested by the last row in the output (`F-statistic: 22.53 ...`)? *Answer:* $H_0 : \beta_1 = \cdots = \beta = 7$ *versus* $H_1 :$ *at least one* $\beta_j \neq 0$, $j = 1, \ldots 7$.

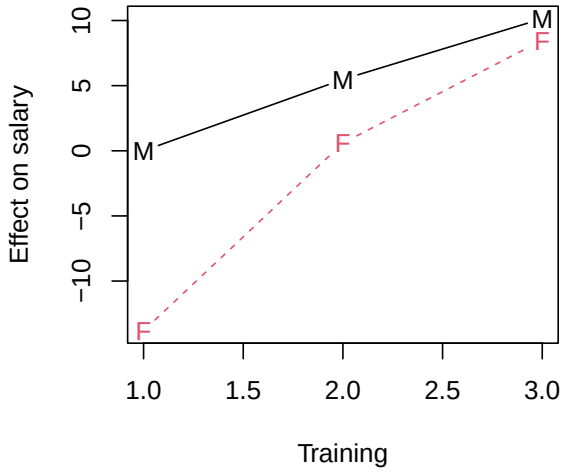
(b) (3 points) Is the interaction significant at the .05 level? State the appropriate null and alternative hypotheses, P-value, and decision. Hint: use the `drop1` output. *Answer:* $H_0 : \beta_6 = \beta_7 = 0$ *versus* $H_1 : \beta_6 \neq 0$ *and/or* $\beta_7 \neq 0$. *Hint: use the* F *test. we find* $P = .007186 < .05$ *so reject* H_0 *and conclude that the effect of training level on salary differs between males and females.*

(c) (3 points) I deleted the value of the F statistic from the output in the previous part. What is the correct value? *Answer:* $F = (396.54/2)/(2337.7/63) \approx (396.54/2)/6.091^2 \approx 5.34$

(d) (3 points) Write one sentence to interpret the `female` estimate ($\hat{\beta}_3 = -13.8068$). *Answer: After controlling for the effects age age and experience, females with low training make $13,806.8 less than males with low training, on the average.*

(e) (5 points) Make an interaction plot showing training level on the horizontal axis and different lines for men and women. Hint: holding age and experience constant, let the intercept for males be the value a, then find the other five predicted values in terms of a, i.e., fill out the table below. For this exercise you may round everything to one decimal place.

Female	Training	$\hat{y}$
0	Low	a
0	Medium	$a + 5.48$
0	High	$a + 10.14$
1	Low	$a - 13.81$
1	Medium	$a - 13.81 + 5.48 + 8.93 = a - 0.6$
1	High	$a - 13.81 + 10.14 + 12.15 = a + 8.48$



5. (12 points) This question investigates how participation in a social media contest affects future spending. A company sponsored a social media contest. Customers in the company's database were invited to write about their relationship with the company on a social media forum. Those who participated by writing at least one word on the forum received a reward worth approximately \$1, and the dummy variable `tx` indicates whether or not a customer participated. In total, 7089 customers participated, and there is a matched control group of 7089 consistent of customers who did not participate, but had similar purchase activities prior to the contest. The total sample size is thus $2 \times 7089 = 14,178$. The variable `y` records the amount spent by each customer in the week following the contest. The variable `x` gives the amount spent per week prior to the contest and will be used as a control variable to account for differences in customer loyalty. Finally, the `wc` variable gives the word count of the entries, where `wc` $= 0$ for all who did not participate. Word count measures *cognitive elaboration*. Note that `tx` $=$ (`wc` > 0). Two models are estimated below.

Model 1

```
Call: lm(log(y + 1) ~ log(x + 1) + tx, data=dat)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.31705    0.04155  -7.631 2.47e-14 ***
log(x + 1)   0.80318    0.01205  66.657  < 2e-16 ***
tx           0.24438    0.02845   8.591  < 2e-16 ***

Residual standard error: 1.693 on 14175 degrees of freedom
Multiple R-squared:  0.2406,Adjusted R-squared:  0.2405
F-statistic:  2246 on 2 and 14175 DF,  p-value: < 2.2e-16
```

Model 2

```
Call: lm(log(y + 1) ~ log(x + 1) + tx + log(wc + 1), data=dat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.30823    0.04166  -7.398 1.46e-13 ***
log(x + 1)   0.80026    0.01209  66.168  < 2e-16 ***
tx           0.05039    0.07657   0.658  0.51053
log(wc + 1)  0.07382    0.02706   2.729  0.00637 **

Residual standard error: 1.693 on 14174 degrees of freedom
Multiple R-squared:  0.241,Adjusted R-squared:  0.2409
F-statistic:  1500 on 3 and 14174 DF,  p-value: < 2.2e-16
```

Use the following notation in answering the questions:

$$\log(y+1) = \beta_0 + \beta_1 \log(x+1) + \beta_2 \texttt{tx} + \beta_3 \log(\texttt{wc}+1) + e,$$

where β_3 is constrained to be 0 in Model 1 and log is the natural log.

(a) (4 points) Based on Model 1, does participation have a significant effect on future spending? Explain. Note: to receive full credit you should state null and alternative hypotheses and do something to determine whether H_0 can be rejected at the 5% level. *Answer: $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$. The P-value is less than 5%, so reject H_0.*

(b) (3 points) Why is the magnitude of the `tx` variable so different in Model 2 (0.050) than in Model 1 (0.244)? *Answer: Multicollinearity. We expect there to be a positive correlation between `tx` and `log(wc+1)` since `tx=0` implies `wc=0`. The effect of `tx` in the Model 1 is overstated.*

(c) (3 points) Now consider Model 2. How do the results from Model 2 change your conclusions about how participation affects future spending. I am looking

for you to summarize the key learnings from Model 2 succinctly. *Answer: The* `tx` *variable is not significant in Model 2, but the word count is. This indicates that participation by itself is not important, and it is the amount of cognitive elaboration that affects future spending.*

(d) (2 points) What do the results of this analysis suggest the company should do in the future when designing social media contests? *Answer: The positive association between word count and future spending suggest that they should do things to encourage more cognitive elaboration in future contests. Students do not need to say this, but regression establishes that there is a relationship, but not causation; there could be other factors that cause someone to elaborate more and increase their spending.*

(e) Use Model 1 for this part. Disregarding the offset of +1, is spend in the week following the contest proportional to spend in the week before the contest? Hint: if we consider unlogged dollars, is $y + 1$ proportional to $x + 1$? Do something to test this and describe the relationship between $y + 1$ and $x + 1$. *Answer: Test* $H_0 : \beta_1 = 1$ *versus* $H_1 : \beta_1 \neq 1$. *A 95% CI is* $0.80318 \pm 1.96 \times 0.0125 = (0.78, 0.83)$, *which does not contain 1. Conclude that* $\beta_1 < 1$, *so we have diminishing returns.*

(f) Use Model 1 for this part and disregard the offset (+1). After controlling for previous week spending, By what percentage does unlogged spending increase with participation? (For example, if some would spend $1000 without participation and $1100 with participation the increase would be 10%.) *Answer:* $e^{0.24438} \approx 27.7\%$ *increase*