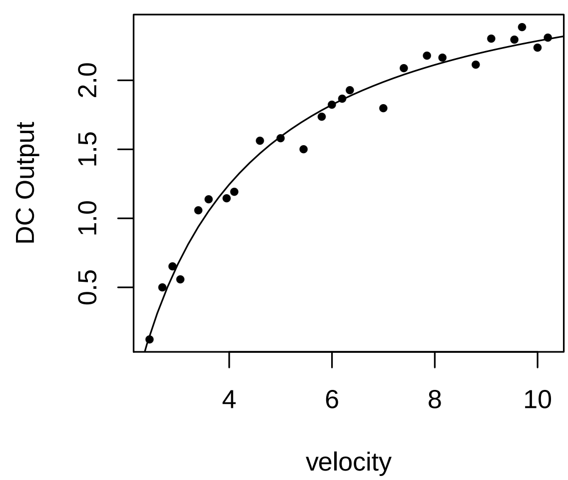
MSiA 401: Predictive Analytics I
Professor E.C. Malthouse
Midterm 1, Fall, 2022

- You may use pencil or pen, a scientific calculator and one sheet of notes (8.5 by 11, both sides). Students may use blank scratch paper.
- The midterm is due by *4:20pm* (80 minutes).

1. (15 points) An engineer investigated using a windmill to generate electricity and collected data on the DC output, y, and the corresponding wind velocity (mph), x. The data are plotted to the right. Theories from windmill operations postulate that DC output approaches an upper limit of approximately 2.5 as wind velocity increases. The engineer fits the following "inverse" model:

$$y = \beta_0 + \beta_1 \left(\frac{1}{x}\right) + e$$



R output is below, and the fitted curve has been superimposed on the plot.

```
Call: lm(formula = output ~ I(1/velocity), data = wind)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.9789     0.0449   66.34   <2e-16 ***
I(1/velocity)   -6.9345     0.2064  -33.59   <2e-16 ***

Residual standard error: 0.09417 on 23 degrees of freedom
Multiple R-squared:   0.98, Adjusted R-squared:  0.9792
```

(a) (2 points) Based on the scatterplot, is the constant variance assumption (homoscedastic) in question? Explain briefly.

(b) (2 points) What does the value of $R^2 = .98$ tell you?

(c) (2 points) What does the value of the residual standard error (0.09417) tell you?

(d) (5 points) Test whether the theoretical limit of 2.5 is plausible, given the data from this experiment. State the null and alternative and do something to decide whether the null can be rejected at the 5% level. Note that the .975 percentile of a t distribution with 23 degrees of freedom is 2.07.

(e) (4 points) Based on inspection of the scatterplot, the engineer also considered the quadratic model to address the nonlinearity: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$. Discuss briefly whether quadratic model would be appropriate for this problem, considering the theories of windmill operation mentioned above.

2. (23 points) This problem examines data from a field experiment on a frequently purchased brand. There are 76 weeks of data on a control panel[1] of families and a matched experimental panel from the same city. Both panels are exposed to the same base level of advertising for the first 52 weeks. For the last 24 weeks, the control panel is continued to be exposed to the base level while the experimental panel is exposed to twice the level of advertising as the control panel. The dependent variable is weekly sales of the brand undergoing the experiment. This is obtained by aggregating over all families that purchase the brand during a given week. The independent variables are the brand's own price and the chief competitive brand's price. A variable `dummy` was created to take the value 0 for the first 52 weeks and 1 for the next 24 weeks. The variable `dummy` can be considered a pre-post variable, i.e., a variable that measures change in the last 24 weeks relative to the first 52 weeks. The variables are as follows:

- `SalesE`: sales of our brand in experimental group
- `SalesC`: sales of our brand in control group
- `BPriceE`: our brand's price in experimental group
- `BPriceC`: our brand's price in control group
- `CPriceE`: competitive brand's price in experimental group
- `CPriceC`: competitive brand's price in control group

(a) (3 points) The following output is from a regression of sales in the experimental group (`SalesE`) on the brand's price (`BPriceE`) in the experimental group, the competitor's price (`CPriceE`) in the experimental group, and `dummy`. Use the following symbols for the true parameters:

$$\text{SalesE} = \beta_0 + \beta_1 \text{BPriceE} + \beta_2 \text{CPriceE} + \beta_3 \text{dummy} + \epsilon$$

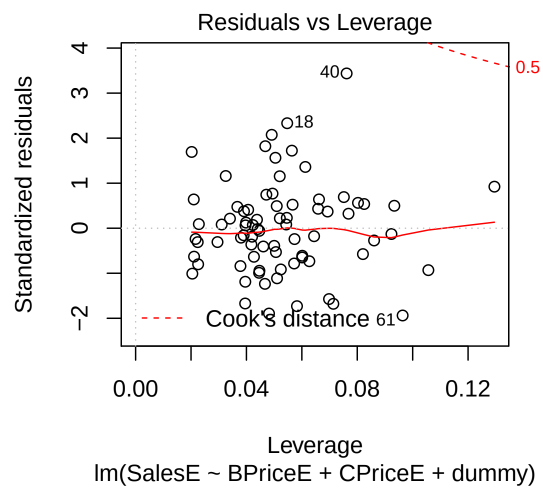
```
Call: lm(formula = SalesE ~ BPriceE + CPriceE + dummy, data = dat)

              Estimate Std. Error t value Pr(>|t|)
(Intercept)   15794.44    2788.30   5.665 2.84e-07 ***
BPriceE      -14541.01    1827.10  -7.958 1.83e-11 ***
CPriceE        7587.98    1845.95   4.111 0.000103 ***
dummy            45.48     600.21   0.076 0.939806
```

[1] Note that a *panel* is a group of people (or sampling units) measured over time.

```
Residual standard error: 2405 on 72 degrees of freedom
Multiple R-squared:  0.4982,Adjusted R-squared:  0.4773
F-statistic: 23.83 on 3 and 72 DF,  p-value: 8.076e-11
```

Test the overall significance of the model us the 5% level of significance. State the null an alternative hypothesis, give the P value, and state your conclusion.

(b) (2 points) The plot to the right is from the standard diagnostics provided by R. What do you conclude from it?



(c) (2 points) The variance inflation factors are as follows:

```
> vif(fit)
 BPriceE  CPriceE    dummy
1.060709 1.040230 1.022822
```

What do you conclude from looking at them?

(d) (2 points) State the literal interpretation of the coefficient for `dummy`. Be precise.

(e) (3 points) Is there evidence to conclude that the value of `dummy` coefficient is different from 0? State the null and alternative hypothesis, P-value, and your conclusion.

(f) (3 points) In managerial terms, what does the hypothesis test on `dummy` indicate about the effectiveness of doubling advertising?

(g) (3 points) The following output is from a regression of sales <u>in the control group</u> (`SalesC`) on the brand's price in the control group (`BPriceC`), the competitor's price in the control group (`CPriceC`), and `dummy`.

```
Call: lm(formula = SalesC ~ BPriceC + CPriceC + dummy, data = dat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19050.3     2859.7   6.662 4.62e-09 ***
BPriceC     -15475.9     1855.0  -8.343 3.51e-12 ***
CPriceC       5989.5     1962.0   3.053  0.00318 **
dummy        -1170.5      571.9  -2.047  0.04434 *
```

Is there evidence to conclude that the value of `dummy` is different from 0? State the null and alternative hypothesis, P-value, and your conclusion.

(h) (5 points) What do you make of the estimate of the `dummy` coefficient (-1170.5) in this regression? What do you think is going on? Can you put the results from the experimental and control groups together to reach a conclusion about the effectiveness of the increased advertising in the experimental panel?

3. (33 points)Short answer and explain briefly.

(a) (3 points) True or false: when the errors of a regression model are heteroscedastic, the OLS estimates will be unbiased.

(b) (3 points) True or false: when the errors of a regression model are uncorrelated and homoscedastic but not normal, the OLS estimates will be unbiased but not BLUE.

(c) (4 points) We estimate the model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, but the real model is $y = \beta_0 + \beta_1 x + 4w + \epsilon$ where w is a variable we didn't think to measure. Will the OLS estimate $\hat{\beta}_1$ underestimate, overestimate, or provide an unbiased estimate of the true β_1. Explain.

(d) (3 points) A manufacturing company is interested knowing the number of hours required (Y) to process an order of a given size (x), and so an analyst regresses Y on x. The company receives an order for 100 units and management wants a prediction of hours with a measure of the uncertainty of the prediction. How do you suggest quantifying the uncertainty? Circle the **best** answer.

i. Confidence interval (CI) for $\hat{\beta}_1$
ii. CI for the mean prediction
iii. Prediction interval
iv. CI for S_e^2

(e) (5 points) A data scientist sees heteroscedasticity in a scatterplot with the spread in dependent variable y increasing with its mean. To address the problem, she considers two variance-stabilizing transformations: $\sqrt{y} = \boldsymbol{\beta}^\mathsf{T}\mathbf{x} + \epsilon$ and $\log y = \boldsymbol{\beta}^\mathsf{T}\mathbf{x} + \epsilon$. Should she compare R^2 values to decide between the square root and log? If not, how?

(f) (3 points) You estimate the model $\widehat{\log(y)} = 1.3 + 0.72 \log(x)$ from a very large sample (so standard errors are near 0). What does your analysis reveal about the relationship between unlogged y and unlogged x?

(g) (8 points) You estimate the model $y = \alpha + \beta x + e$ with OLS, where both x and y are measured in centimeters. Suppose that your colleague also analyzed the data, but expressed the predictor variable in millimeters (i.e., colleague regressed y on $x^* = 10x$). How will the following will change in your colleague's model versus yours?

i. Slope estimate $\hat{\beta}$

ii. The t statistic for $\hat{\beta}$: $\hat{\beta}/\text{SE}(\hat{\beta})$

iii. Mean squared error S_e^2

iv. R^2

(h) (4 points) You have been hired as a consultant to answer the following question for a fertilizer company: does anti-fungal soil treatment reduce the height of a certain plant of interest? Past experience suggests that fungus reduces plant growth, which is why the company is investigating the anti-fungal soil treatment. To answer the question, a large number of plants was seeded and sprout in a greenhouse. The sprouts selected all have (roughly) the same initial height, and were randomly assigned to different treatment conditions, e.g., no anti-fungal soil treatment or soil that has been treated with a fixed amount of anti-fungal treatment. The sprouts were given the same amount of water and sunlight, and isolated from each other. After a period of time, two variables were measured, the amount of fungus on the plant and the height. To answer the question of whether treatment affects height, you regress height on treatment. Should you also include the amount of fungus as a control variable in your regression? Explain briefly.

4. (4 points) A cable TV service provider offers three services: video, phone and data (internet). A data scientist found the following segments. For example, Data basic only pays for one service, data ($35/month on average). Data/Phone pays for two services: data ($40/month) and phone ($25/month). The three "Triple" segments are "triple plays" that pay for all three services. The analyst notes that the more services that a customer has the greater the customer's monthly retention rate (probability of remaining a customer each month). The analyst concludes that more services imply more loyalty, and therefore the company should cross-sell additional services to increase retention. Is this reasoning sound? (Hint: do not focus on dollar amounts.)

Segment	Number of Services	Video Mean	Data Mean	Phone Mean	Retention Rate
Data basic	1	$0	$35	$0	97.1%
Data star	1	$0	$60	$0	97.0%
Data+Video basic	2	$65	$40	$0	97.1%
Data+Phone	2	$0	$40	$25	98.4%
Data+Video star	2	$100	$50	$0	97.7%
Triple play basic	3	$75	$35	$25	98.8%
Triple play	3	$110	$40	$20	98.9%
Triple play star	3	$150	$45	$15	99.0%

5. (22 points) The density of a finished product (y) is an important performance characteristic (higher is better). It can be controlled by three main manufacturing variables, and our interest is in understanding their effects on density:

- x_1 = the amount of water in the product mix,
- x_2 = the amount of reworked material in the product mix,

- x_3 = the temperature of the mix

A designed experiment with 48 batches was run to assess the impact of the predictor variables on density, where x_1–x_3 were manipulated experimental variables with an orthogonal design. Variables x_1–x_3 each have two levels (low or high), and I have represented them as **dummy variables** in the regressions, i.e., $x_j = 1$ for high and $x_j = 0$ for low, $j = 1, 2, 3$. We also measured two control variables during the process:

- x_4 = the air temperature in the drying chamber, which is affected by the temperature of the mix and the heat generated by the process.
- x_5 = temperature rise (increase) during the process.

I have provided a correlation matrix and output from three models on next page. For example, according to Model 1 the estimate for `x1(H)` is for the high value of x_1, so density is 3.77 greater when x_1 is high versus low, on average.

(a) (3 points) $\text{Corr}(x_1, y) = 0.52$ is very highly significant ($P = 0.00017$) yet the slope for x_1 in Model 1 is not significant ($P = 0.174$). Why do you think this happens?

(b) (2 points) Using Model 1 and the correlation matrix, what do you conclude about the effect of x_2 on y?

(c) (4 points) Using Model 2, how, if at all, does x_2 affect y?

(d) (5 points) Use Model 2 for this part. How, if at all, does x_1 affect y? (The interactions are already significant and I'm not looking for any further significance tests—just interpret parameter estimates to answer the question.)

(e) (5 points) Which of the three models do you suggest and why? You will receive 1 point for the right model, and 4 points for your rationale. Think about your answer before you start to write and prioritize the most important point(s).

(f) (3 points) Use Model 2 for this part. Among these treatments, which values of x_1–x_3 maximize density?

```
      x1    x2   x3    x4    x5
x1  1.00  0.00 0.00 -0.45 -0.23
x2  0.00  1.00 0.00 -0.05 -0.08
x3  0.00  0.00 1.00  0.64  0.64
x4 -0.45 -0.05 0.64  1.00  0.96
x5 -0.23 -0.08 0.64  0.96  1.00
y   0.52 -0.11 0.32 -0.33 -0.27
```

```
===============  Model 1  ===============
                   Std.      t
      Estimate   Error  value  Pr(>|t|)
(Int) 237.5347 49.0263  4.845 1.76e-05 ***
x1(H)   3.7743  2.7285  1.383    0.174
x2(H)  -2.3369  1.4528 -1.609    0.115
x3(H)  12.5138  2.1388  5.851 6.51e-07 ***
x4     -0.1538  0.2074 -0.742    0.462
x5     -3.8043  5.8717 -0.648    0.521
Residual standard error: 4.96 on 42 DF
Mult R-sq:  0.6375, Adj R-sq:  0.5944
F-statistic: 14.77 on 5, 42 DF, p=2.322e-08
```

```
===================  Model 2  ===================
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  100.708      1.443  69.796  < 2e-16 ***
x1(H)          6.625      2.041   3.247   0.0023 **
x2(H)         -8.917      1.666  -5.352 3.38e-06 ***
x3(H)         10.917      1.666   6.552 6.38e-08 ***
x1(H):x2(H)   14.583      2.356   6.189 2.12e-07 ***
x1(H):x3(H)  -11.917      2.356  -5.057 8.83e-06 ***
Residual standard error: 4.081 on 42 degrees of freedom
Multiple R-squared:  0.7546,Adjusted R-squared:  0.7254
F-statistic: 25.83 on 5 and 42 DF,  p-value: 8.162e-12
```

```
===================  Model 3  ===================
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -144.5638   145.8280  -0.991  0.32748
x1(H)         21.1811     9.0373   2.344  0.02415 *
x2(H)        -12.0084     2.4135  -4.976 1.28e-05 ***
x3(H)          9.4110     1.8775   5.012 1.14e-05 ***
x4             0.5050     0.3462   1.459  0.15246
x5            -3.6107     5.6998  -0.633  0.53003
x1(H):x2(H)   22.0194     4.8722   4.519 5.40e-05 ***
x1(H):x3(H)  -31.0186    11.1844  -2.773  0.00839 **
Residual standard error: 4.03 on 40 degrees of freedom
Multiple R-squared:  0.7721,Adjusted R-squared:  0.7323
F-statistic: 19.36 on 7 and 40 DF,  p-value: 5.127e-11
```

6. There will also be something similar to HW4 about `drop1` and the F test.