Name ____

Fall 2021 **STAT 587-2** **Final exam**

(50 pts)

Instructions:

1. Full credit will be given only if you show your work.
2. The questions are not necessarily ordered from easiest to hardest.
3. You are allowed to use any resource except aid from another individual.
4. Aid from another individual will automatically earn you a 0.

1. Use the scatterplots on the page titled "Scatterplots" to answer the following questions. We will consider 3 possible regression models:

   - no transformation needed (x-y)
   - log explanatory variable only (log(x)-y)
   - log response variable only (x-log(y))

   For each plot, indicate which of the following models will fit the data best using the code in parentheses, e.g. x-log(y). If none of these models will fit the data well, then write "none". (1 pt each)

   - a
     Answer: x-y
   - b
     Answer: x-log(y)
   - c
     Answer: log(x)-y
   - d
     Answer: x-y
   - e
     Answer: log(x)-y
   - f
     Answer: x-log(y)
   - g
     Answer: x-y
   - h
     Answer: x-log(y)
   - i
     Answer: none (x^2)
   - j
     Answer: x-y

2. Use the code on the page titled "Food's Carbon Footprint" to answer the following questions.

   (a) How many observations are in the data set? (1 pt)

   Answer:

```
m$df.residual+length(coef(m))
## [1] 650
```

   (b) How many levels of food category are in the data set? (1 pt)

   Answer: From the table, there are 5: Beef, Eggs, Fish, Pork, and Poultry

   (c) What is the reference level for food category? (1 pt)

   Answer: Beef

   (d) If the design is balanced (with respect to food categories), how many countries are represented in the data set? (1 pt)

   Answer: Since there are 650 observations and 5 food categories, then there should be $650/5 = 130$ countries.

   (e) What is the estimate for the residual standard deviation? (1 pt)

   Answer:

```
summary(m)$sigma
## [1] 0.0029126
```

   (f) Provide a point estimate for the CO2 emissions for Fish when consumption is 10 kg/person/year. Show your work for full credit. (5 pts)

   Answer: Calculate

$$308.5787635 + 10 \times 30.8579218 + -292.6121981 + 10 \times -29.2612423 = 31.9333612$$

```
coef(m)[1]+10*coef(m)[2]+coef(m)[4]+10*coef(m)[8]
## (Intercept)
##    31.93336
```

(g) Provide interpretations for the following values: (2 pts each)

- 308.5788 (Intercept)
  Answer: The mean CO2 emissions across countries associated with beef when consumption is zero is estimated to be 309 kg/person/year.

- 30.8579 I(consumption-10)
  Answer: Each additional kg/person/year increase in beef consumption is associated with an additional 31 kg/person/year of CO2 emissions.

- -299.3924 categoryEggs
  Answer: When consumption is 10, eggs produce 299 less kg CO2/person/year than beef.

(h) Explain why R^2 is 1. (2 pts)

Answer: The model fits the data (almost) perfectly. This is likely due to assumptions about the amount of CO2 emissions per consumption across the different food categories.

(i) Calculate a 95% confidence/credible interval for the coefficient for the indicator for Eggs. (2 pts)

Answer:

```
s <- summary(m)$coefficients[3,]
s[1] + c(-1,1) * qt(0.975, df = m$df.residual) * s[2]
## [1] -299.3931 -299.3916
```

3. Use the data file `SRM_1540e_Fibrous_Glass_Board.csv` to answer the following questions. In this dataset, we evaluate the thermal conductivity (W/mK) of SRM 1540e Fibrous Glass Board at various temperatures (K) and air pressure (kPa).

Answer:

```
d <- read_csv("SRM_1540e_Fibrous_Glass_Board.csv",
              col_types = cols(.default = col_double()))
```

(a) What values of temperature exist in the experiment? (2 pts)

Answer:

```
unique(d$temperature)
## [1] 280 320 340 300 360
```

(b) Is the design of the experiment complete? Why or why not? (2 pts)

Answer:

```
d %>% group_by(temperature, air_pressure) %>% summarize(n = n())

## `summarise()` has grouped output by 'temperature'.  You can override using
the `.groups` argument.

## # A tibble: 15 x 3
## # Groups:   temperature [5]
##    temperature air_pressure     n
##          <dbl>        <dbl> <int>
##  1         280           60     3
##  2         280           80     3
##  3         280          100     3
##  4         300           60     3
##  5         300           80     3
##  6         300          100     3
##  7         320           60     3
##  8         320           80     3
##  9         320          100     3
## 10         340           60     3
## 11         340           80     3
## 12         340          100     3
## 13         360           60     3
## 14         360           80     3
## 15         360          100     3
```

Yes because every combination of temperature (280, 300, 320, 340, and 360) and air pressure (60, 80, 100) exist.

(c) Is the design of the experiment balanced? Why or why not? (2 pts)

Answer: Yes, because every combination of temperature and air pressure has the same number of observations.

(d) Is this a replicated experiment? Why or why not? (2 pts)

Answer: Yes, because every combination of temperature and air pressure has 3 replicates.

(e) What is the correlation between temperature and air pressure in these data? Explain why this is the correlation. (2 pts)

Answer:

```
cor(d$temperature, d$air_pressure)
## [1] 0
```

We knew this would be 0 since the design is complete and balanced.

Important: Throughout this question, treat temperature and air pressure as categorical (rather than continuous) variables.

(f) Conduct an F-test to determine whether an interaction between temperature and air pressure is needed.

Answer:

```
m <- lm(thermal_conductivity ~ factor(temperature) * factor(air_pressure), data = d)
```

- F-statistic (1 pt)

  Answer:

  ```
  drop1(m, test="F")$`F value`[2]
  ## [1] 0.02421907
  ```

- p-value (1 pt)

  Answer:

  ```
  drop1(m, test="F")$`Pr(>F)`[2]
  ## [1] 0.9999952
  ```

- conclusion (1 pt)

  Answer: There is no evidence that the interaction needs to be included.

(g) With the **additive model**, provide a prediction with uncertainty for the thermal conductivity at temperature of 320 K and air pressure of 100 kPa.

- point estimate (1 pt)

  Answer:

  ```
  m <- lm(thermal_conductivity ~ factor(temperature) + factor(air_pressure), data = d
  p <- predict(m, newdata = data.frame(temperature = 320, air_pressure = 100),
                                        interval = "prediction", level = 0.9)
  p[1]
  ## [1] 0.03510448
  ```

- 90% Prediction interval (2 pts)

  Answer:

  ```
  p[2:3]
  ## [1] 0.03444228 0.03576669
  ```

(h) With the **additive model**, calculate a contrast to compare the difference in thermal conductivity between temperatures 360 and 280 averaged over air pressure.

- point estimate (2 pts)

  Answer:

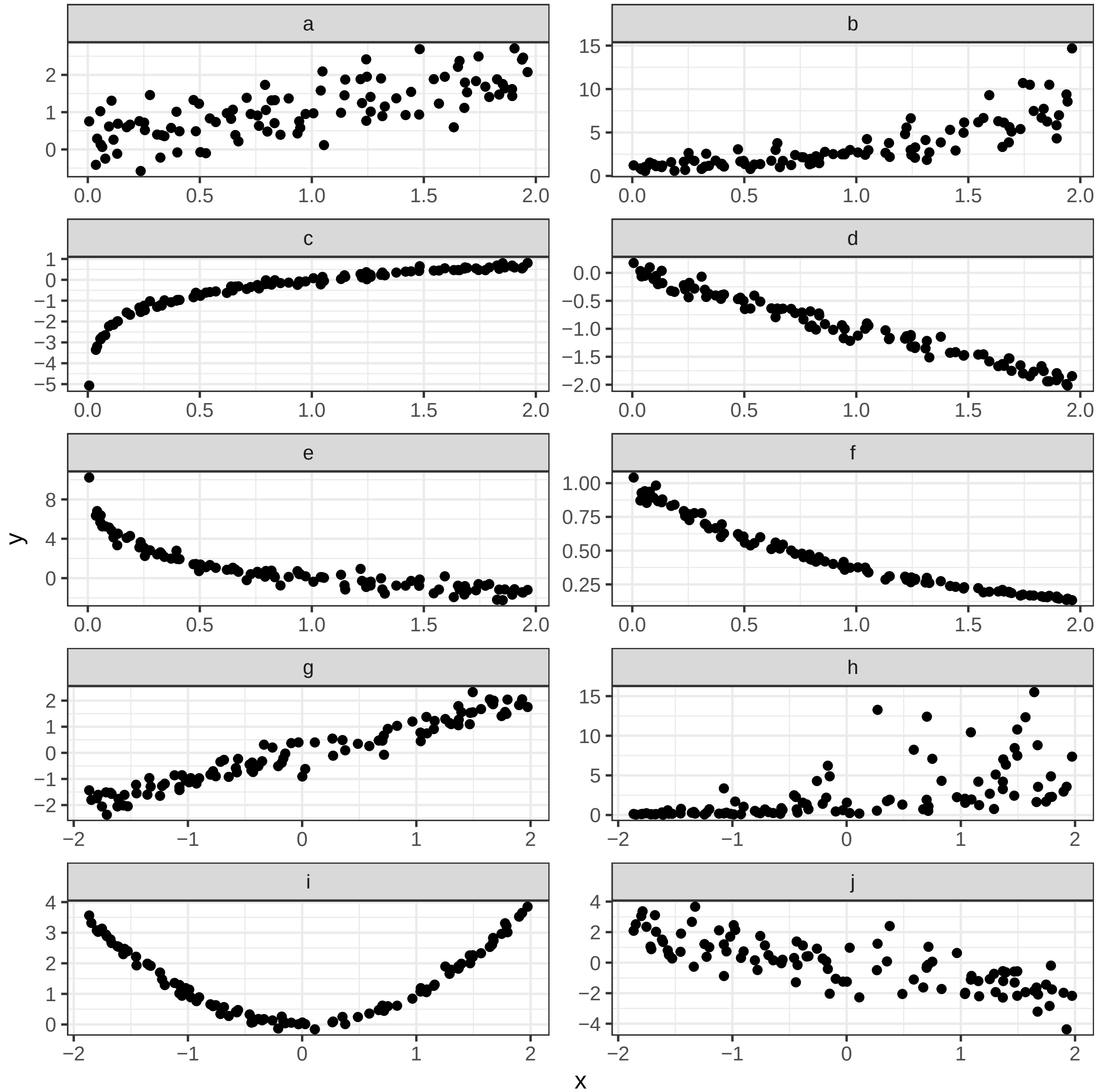
  ```
  em <- emmeans(m, pairwise ~ temperature)
  co <- as.data.frame(confint(em, level = 0.99, adjust = "none")$contrasts)
  co[co$contrast == "280 - 360", "estimate"]
  ## [1] -0.008793844
  ```

- 99% Confidence/credible interval (2 pts)

  Answer:

  ```
  co[co$contrast == "280 - 360", c("lower.CL","upper.CL")]
  ##       lower.CL     upper.CL
  ## 4 -0.009260894 -0.008326795
  ```

Scatterplots



Food's Carbon Footprint

The following regression model analyzes the relationship between carbon dioxide emissions (kg CO2/person/year) and food consumption (kg/person/year) across a large number of countries and food categories.

```
table(food_consumption$category)
##
##    Beef    Eggs    Fish    Pork Poultry
##     130     130     130     130     130
m <- lm(co2_emission ~ I(consumption-10) * category, data = food_consumption)
```

```
summary(m)
##
## Call:
## lm(formula = co2_emission ~ I(consumption - 10) * category, data = food_consumption)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0053348 -0.0024859 -0.0003105  0.0024896  0.0053201
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          3.086e+02  2.616e-04 1179562   <2e-16 ***
## I(consumption - 10)                  3.086e+01  2.656e-05 1161701   <2e-16 ***
## categoryEggs                        -2.994e+02  3.774e-04 -793256   <2e-16 ***
## categoryFish                        -2.926e+02  3.786e-04 -772841   <2e-16 ***
## categoryPork                        -2.732e+02  3.790e-04 -720716   <2e-16 ***
## categoryPoultry                     -2.978e+02  4.155e-04 -716745   <2e-16 ***
## I(consumption - 10):categoryEggs    -2.994e+01  5.746e-05 -521012   <2e-16 ***
## I(consumption - 10):categoryFish    -2.926e+01  2.979e-05 -982228   <2e-16 ***
## I(consumption - 10):categoryPork    -2.732e+01  3.117e-05 -876342   <2e-16 ***
## I(consumption - 10):categoryPoultry -2.978e+01  3.186e-05 -934733   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002913 on 640 degrees of freedom
## Multiple R-squared:      1,Adjusted R-squared:      1
## F-statistic: 3.219e+11 on 9 and 640 DF,  p-value: < 2.2e-16
```

```
anova(m)
## Analysis of Variance Table
##
## Response: co2_emission
##                               Df   Sum Sq Mean Sq      F value    Pr(>F)
## I(consumption - 10)            1  1717832 1717832 202497070305 < 2.2e-16 ***
## category                       4 13733627 3433407 404728154460 < 2.2e-16 ***
## I(consumption - 10):category   4  9121848 2280462 268819640074 < 2.2e-16 ***
## Residuals                    640        0       0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```