

Name _____

Fall 2020

STAT 587-2

Final exam
(50 points)

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.

Short answer (Multiple choice on Canvas)

1. Provide answers to the following questions.

(a) Name two ways to generate random samples.

Answer: Use a random number generator or roll a dice.

(b) How can we ensure our inferences are generalizable to a larger population?

Answer: Randomly sample from the population

(c) How can we make statistically valid cause-and-effect statements?

Answer: Randomly assign treatments to experimental units.

(d) State the 4 simple linear regression model assumptions.

Answer: Errors are 1) normally distributed, 2) independent, and 3) homoskedastic (constant variance); there is a linear relationship between the explanatory variable and the expected response.

Medical Cost Personal Datasets

Please answer questions on the next two pages based on the analysis provided on the page titled “Medical Cost Personal Datasets - Analysis.”

2. Identify the numeric values for the following quantities.

(a) Point estimate for the intercept

[Answer:](#) 7.3980216

(b) Point estimate for the coefficient for age

[Answer:](#) 0.0352984

(c) Point estimate for the coefficient for sex

[Answer:](#) -0.0820905

(d) Point estimate for the coefficient for bmi

[Answer:](#) 0.0013003

(e) Point estimate for the coefficient for the indicator that the individual is a smoker

[Answer:](#) 0.1893146

(f) Point estimate for the coefficient for the interaction between bmi and the indicator that the individual is a smoker.

[Answer:](#) 0.0444494

(g) Point estimate for the error standard deviation

[Answer:](#) 0.4504864

(h) Coefficient of determination

[Answer:](#) 0.7608848

Medical Cost Personal Datasets (cont.)

Calculate the following quantities.

- (i) The number of observations in the data set

[Answer:](#) 1338

- (j) Multiplicative effect of a one year increase in age on median charges

[Answer:](#) 1.0359288

- (k) Multiplicative effect of being male compared to female on median charges

[Answer:](#) 0.9211886

- (l) Multiplicative effect of a one unit increase in bmi on median charges for **non-smokers**

[Answer:](#) 1.0013012

- (m) Multiplicative effect of a one unit increase in bmi on median charges for **smokers**

[Answer:](#) 1.0468124

- (n) Point estimate for the error variance

[Answer:](#) 0.202938

Model diagnostics

3. For the following plots, indicate what simple linear regression model assumptions may be evaluated. Example plots are provided on the page titled “Model diagnostics - plots”. Note: this question is **not** asking what assumptions are violated, but rather what assumptions **may** be evaluated.

(a) Residual Plot

Answer: Errors have constant variance and relationship between explanatory variable and response is linear.

(b) Q-Q Plot

Answer: Errors are normally distributed.

(c) COOK’s D Plot

Answer: No model assumptions can be evaluated using this plot.

(d) Index Plot

Answer: Errors are independent.

(e) Location-Scale Plot

Answer: Errors have constant variance.

Methane Production

4. An experiment was conducted with a collection of biogas digesters to determine the effect of month (as a proxy for temperature) and insulation on methane production (L/kg volatile solids [VS]). Analyze the data in `methane.csv` and write a paragraph summarizing your results. In your summary, include
- complete sentences (1 point)
 - a p -value for a test of the interaction between month and insulation and a conclusion (whether the interaction should be included in the model) (2 points)
 - an estimate and 95% prediction interval for methane production in an insulated digester in Dec (3 points)
 - an estimate and 95% confidence interval for the contrast of the effect of insulation averaged over month (3 points)
 - a statement about whether model assumptions appear to be met based on diagnostic plots (1 point)
 - a statement about any other model assumptions you may be concerned about that cannot be addressed by the diagnostic plots (1 point)

You may include code (make it clear) for partial credit.

[Answer:](#)

```
d = read.csv("methane.csv")
m = lm(methane ~ month*insulation, data = d)
dr = drop1(m, test = "F")
# resid_panel(m, plot = my_plots)
p = predict(m, data.frame(month = "Dec", insulation = "Yes"),
            interval = "prediction") %>% round(1)
e = emmeans(m, pairwise~insulation) %>% confint
```

NOTE: Results may be misleading due to involvement in interactions

An experiment was conducted with a collection of biogas digesters to determine the effect of month (as a proxy for temperature) and insulation on methane production. A test for the interaction of month and insulation suggests evidence against the non-interaction model ($p=0.038$) and diagnostic checks indicated no clear model violations, thus the interaction was included. A new insulated digester in December is expected to have a methane production of 209.6 with a 95% prediction interval of (207.2, 212.1). An insulated digester is expected to produce 5.16 (4.16, 6.16) L/kg VW more methane averaging over the months of Oct, Nov, Dec, and Jan. Since multiple measurements are taken on each digester, the observations are likely to not be independent.

COVID-19 Filters

5. A company is developing filters to eliminate SARS-CoV-2, the coronavirus that causes COVID-19. The file `filters.csv` contains data from an experiment conducted to study the effectiveness of different filter designs in removing inactivated virus particles. While running the experiment, the researchers recorded the pressure drop (Pa) and the percentage of virus particles removed. In practice, the filters will be run at a pressure of 50 Pa.

Fit the model with an interaction between pressure and design. Provide

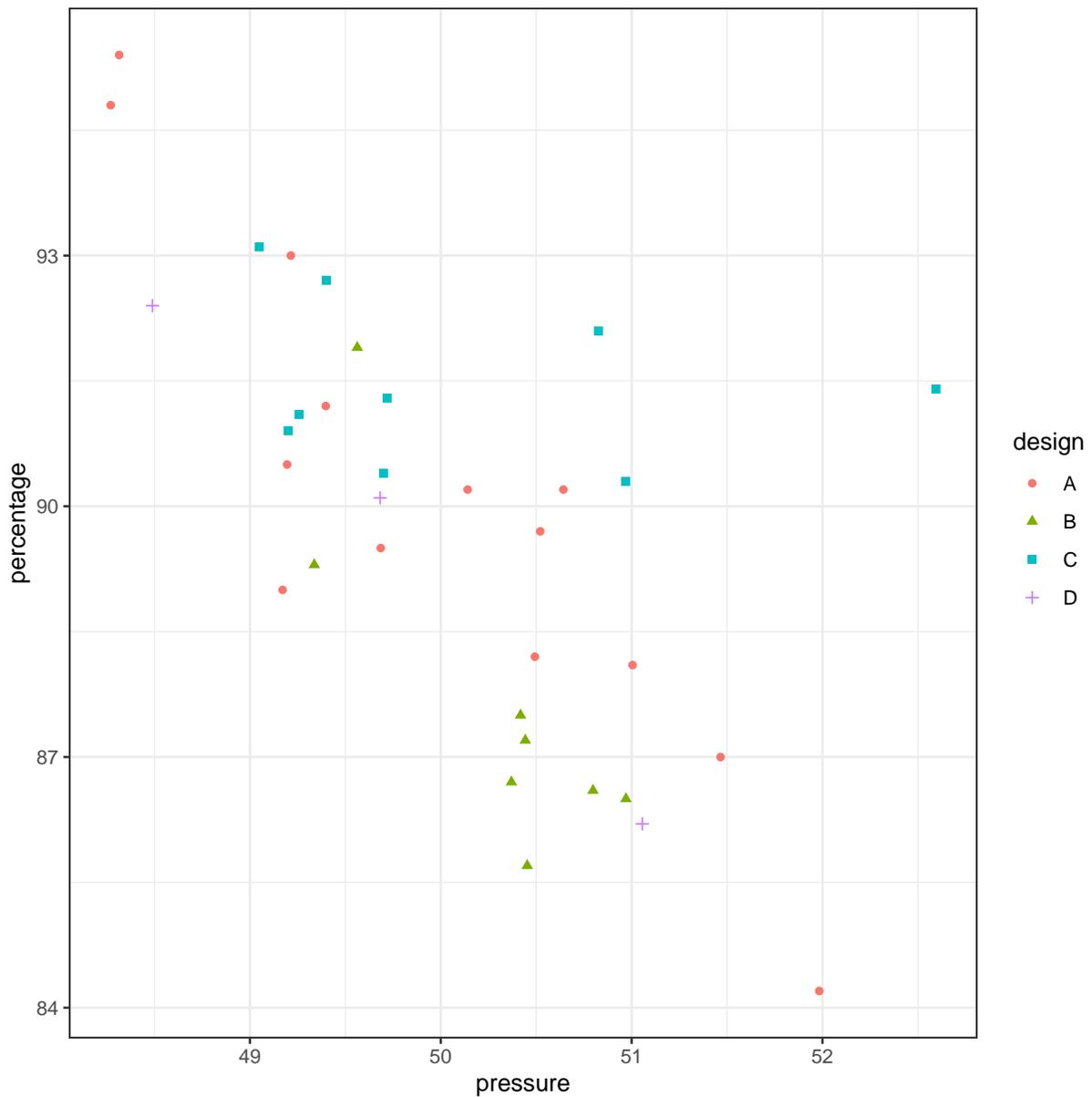
- a plot of the data containing
 - the response variable on the y-axis (1 point)
 - the continuous explanatory variable on the x-axis (1 point)
 - the categorical explanatory variable distinguished by color, shape, or both (1 point)
 - x axis label with units, if appropriate (1 point)
 - y axis label with units, if appropriate (1 point)
 - legend (1 point)
- a paragraph summary presenting your findings that includes
 - complete sentences (1 point)
 - an estimate and 95% credible interval for the expected difference in percentage removed between design C and design A at a pressure of 50 Pa (3 points)
 - proportion of variability described by the model (1 point)

You may include code/output (make it clear) for partial credit.

[Answer:](#)

```
d = read.csv("filters.csv")

ggplot(d, aes(pressure, percentage, color = design, shape = design)) +
  geom_point() +
  theme_bw()
```



```

m = lm(percentage ~ pressure*design, data = d)
e = emmeans(m, pairwise ~ design | pressure,
            adjust = "none",
            at = list(pressure = 50)) %>%
  confint

```

An experiment was conducted to study the effectiveness of different filter designs in removing inactivated virus particles. At 50 Pa of pressure, design C removed 1.51% more virus particles with a 95% credible interval of (0.43%, 2.59%). The model containing an interaction between pressure and design accounted for 83% of the variability in the percentage.

Medical Cost Personal Datasets - Analysis

The analysis here uses data from a kaggle competition.

```
summary(m_insurance)
##
## Call:
## lm(formula = log(charges) ~ age + sex + bmi * smoker, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.96123 -0.19875 -0.04082  0.09212  2.16225
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.398022   0.077229  95.793 < 2e-16 ***
## age           0.035298   0.000883  39.974 < 2e-16 ***
## sexmale      -0.082090   0.024769  -3.314 0.000944 ***
## bmi          0.001300   0.002301   0.565 0.572044
## smokeryes    0.189315   0.153099   1.237 0.216473
## bmi:smokeryes 0.044449   0.004891   9.087 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4505 on 1332 degrees of freedom
## Multiple R-squared:  0.7609, Adjusted R-squared:  0.76
## F-statistic: 847.7 on 5 and 1332 DF,  p-value: < 2.2e-16
confint(m_insurance)
##              2.5 %       97.5 %
## (Intercept)   7.246516992  7.549526118
## age           0.033566172  0.037030722
## sexmale      -0.130681129 -0.033499813
## bmi          -0.003213115  0.005813785
## smokeryes    -0.111027621  0.489656903
## bmi:smokeryes 0.034853857  0.054045002
```

Model diagnostics - plots

