Stat 423/504 Winter 2020 Midterm

1. (10 points) We want to study the relationship between motor fuel consumption and related variables in each state in the USA, or in the District of Columbia for the year 2001. Specifically, we want to predict the response `Fuel`(normalized by state population), using the following predictors:

   - `Tax`: Gasoline state tax rate in cents per gallon,
   - `Dlic`: Number of licensed drivers in the state, normalized by state population,
   - `Income`: Per capita person income of year 2000,
   - `log10(Miles)`: (Base 10) logarithm of the miles of Federal-aid highway in the state.

   The following `R` output is a partial summary of the multiple linear model, as well as the partial output of an anova comparison with an empty model.

```
lm(formula = Fuel ~ Tax + Dlic + Income + log10(Miles), data = fuel2001)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  154.1928   194.9062
Tax           -4.2280     2.0301
Dlic           0.4719     0.1285
Income        -6.1353     2.1936
log10(Miles)  61.6061    21.5001
---

Residual standard error: 64.89 on 46 degrees of freedom
>
Analysis of Variance Table

Model 1: Fuel ~ Tax + Dlic + Income + log10(Miles)
Model 2: Fuel ~ 1
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     46 193700
2     50 395694 -4   -201994 11.992 9.331e-07
---
```

   (a) (1 point) How many samples are used to fit the least-squares multiple linear model above?
   Solution: 51.

   (b) (1 point) Compare the average fuel consumption in

   - State 1: $Tax = 10, Dlic = 1000, Income = .05, Miles = 10$
   - State 2: $Tax = 10, Dlic = 1000, Income = .05, Miles = 100$

   Solution: The average fuel consumption in State 2 is larger by 61.6061.

   (c) (1 point) Write the equation of the estimated multiple linear model in the form:

$$\widehat{Fuel} = \widehat{\beta_0} + \widehat{\beta_1} Tax + \cdots$$

   Use the partial summary above to get the estimsted coefficients.

   Solution: $\widehat{Fuel} = 154.19 - 4.23 * \texttt{Tax} + 0.47 * \texttt{Dlic} - 6.14 * \texttt{Income} + 61.61 * \texttt{log(Miles)}$

   (d) (5 points) We are assuming a linear model

$$Fuel_i = \beta_0 + \beta_1 Tax_i + \beta_2 Dlic_i + \beta_3 Income_i + \beta_4 log(Miles)_i + \epsilon_i,$$

   for $i \in \{1, \ldots, n\}$.

i. (2 points) To calculate the p-values given in the R summary, what assumptions do we need to make on the errors ϵ_i?

Solution: We assume that the errors $\underline{\epsilon} = (\epsilon_i, \dots, \epsilon_n)^T$, in our model are multivariate normal with mean vector zero, and variance $\sigma^2 I_n$, where I_n is the identity matrix. Another answer: The errors ϵ_i are i.i.d. univariate normal with mean zero and variance σ^2, $i \in \{1, \dots, n\}$. **(0.5 Points for each correct assumption.)**

ii. (1 point) According to the R output above, what is $\hat{\sigma}^2$ equal to?

Solution: $\hat{\sigma}^2 = 64.89^2$

iii. (1 point) Compute the observed value of the test statistic that tests the the null hypothesis $\beta_{Dlic} = 0$.

Solution: The observed test statistic is:

$$\frac{\hat{\beta_{Dlic}}}{SE(\hat{\beta_{Dlic}})} = \frac{0.4719}{0.1285} \approx 3.672.$$

iv. (1 point) Would you reject the null hypothesis $\beta_{Dlic} = 0$ at the 5% level? Explain your reasoning.

Note: In this exercise you can use the following properties:
- For a random variable $X \sim t_{30}$, $Pr(X < 2.042) = 0.975$.
- For a random variable $X \sim t_{36}$, $Pr(X < 2.028) = 0.975$.
- For a random variable $X \sim t_{40}$, $Pr(X < 2.021) = 0.975$.
- For a random variable $X \sim t_{46}$, $Pr(X < 2.013) = 0.975$.

Solution: Yes, we would reject this null hypothesis at the 5% level. The distribution our test statistic follows under the null is t_{46}, 46 are the residual degrees of freedom (as can be seen from the output). Since our observed test statistic is 3.672 and that is larger than 2.013, the p-value of this test will be lower than 0.05.

(e) (2 points) Now we fit a simple linear regression with response `Fuel` and predictor `Tax`. We print a partial summary of this linear model and also print the 95% confidence interval for the average fuel consumption in a state that has Tax=.5.

```
lm(formula = Fuel ~ Tax, data = fuel2001)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  715.485     55.770  12.829   <2e-16 ***
Tax           -5.078      2.701  -1.881    0.066 .
---
Residual standard error: 86.79 on 49 degrees of freedom

> conf_interval <- predict(lm(Fuel ~ Tax, data = fuel2001),
newdata=data.frame(Tax=.5), interval="confidence",
+                            level = 0.95)
> conf_interval
       fit      lwr      upr
1 712.9461 603.5189 822.3733
```

Suppose that the relevant critical value is $t_{.975,n-2} = 2.009$. What is the 95% prediction interval for the average fuel consumption in a state that has Tax=.5?

Solution: The prediction interval is computed as $\widehat{y^*} \pm t_{.975,n-2} SE(y^* - \widehat{y^*}|Tax = .5)$. The confidence interval is computed as $\widehat{y^*} \pm t_{.975,n-2} SE(\widehat{y^*}|Tax = .5)$, where

$$SE(y^* - \widehat{y^*}|Tax = .5)^2 = SE(\widehat{y^*}|Tax = .5)^2 + \hat{\sigma}^2.$$

THen from the output: $SE(\widehat{y^*}|Tax = .5) = (712.9461 - 603.5189)/2.009 = 54.46849$ and $SE(y^* - \widehat{y^*}|Tax = .5) = \sqrt{54.46849^2 + 86.79^2} = 102.4662$.

So the prediction interval is $712.9461 \pm 2.009 * 102.4662 \approx (507.0915, 918.8007)$.

```
> pred_interval <- predict(lm(Fuel ~ Tax, data = fuel2001),
+                             newdata=data.frame(Tax=.5), interval="prediction",
+                               level = 0.95)
> pred_interval

       fit     lwr      upr
1 712.9461 507.056 918.8362
```

2. (9+1 Bonus points) Austin, an actuary, would like to determine some common explanations for expensive bodily injury claims in car accidents. He wants to model the response `LOSS` (a numeric vector for the amount of money the insurance lost in thousands), using the following predictors:

   - `TIME`: (continuous) time until medical service received in hours
   - `ATTORNEY`: 0 if claimant is represented by an attorney; 1 otherwise
   - `SEATBELT`: 0 if the claimant wore a seatbelt; 1 otherwise

   To begin, Austin fits a linear model with 2-way interaction terms. The following R output is a summary of the fitted model.

```
lm(formula = LOSS ~ (ATTORNEY + TIME + SEATBELT)^2, data = autobi)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         7.0623     0.9420   7.497 1.35e-13
ATTORNEY           -5.3696     1.3482  -3.983 7.26e-05
TIME                0.5056     0.3797   1.331  0.18336
SEATBELT           34.1535     6.4057   5.332 1.18e-07
ATTORNEY:TIME      -0.4171     0.5454  -0.765  0.44460
ATTORNEY:SEATBELT -23.6934     8.3194  -2.848  0.00448
TIME:SEATBELT      -6.3064     2.0468  -3.081  0.00211
---

Residual standard error: 15.65 on 1084 degrees of freedom
Multiple R-squared:  0.06942,Adjusted R-squared:  0.06427
F-statistic: 13.48 on 6 and 1084 DF,  p-value: 8.667e-15
```

   (a) (1 point) Write down the predictor variables significant at the $\alpha = .05$ level according to the individual hypothesis tests above.

   Solution: `SEATBELT` and `ATTORNEY` are significant. Two interaction terms are significant: `ATTORNEY:SEATBELT`, and `TIME:SEATBELT`.

   (b) (1 point) How many different fitted lines are implied by the above model?

   Solution: There are 4 different fitted lines due to 4 possible groups: attorney used and seatbelt worn, attorney used and seatbelt not worn, attorney not used and seatbelt worn, and attorney not used and seatbelt not worn.

   (c) (2 points) Write down all of the different slopes for the fitted lines in the model above.

   Solution:

   - 0.5056
   - 0.5056 + (-0.4171)
   - 0.5056 + (-6.3064)
   - 0.5056 + (-0.4171) + (-6.3064)

(d) (1 point) Using the model above, predict the average LOSS for a 27 year old female driver for whom it took 1 hour to receive medical care, who was represented by an attorney and who was wearing a seatbelt at the time of the accident.

Solution: $7.0623 + 0.5056 = 7.5679$

(e) (1 point) Austin decides to present the following model to his boss Caitlyn.

```
lm(formula = LOSS ~ ATTORNEY + SEATBELT + ATTORNEY * SEATBELT,
    data = autobi)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         7.9548     0.6640  11.980  < 2e-16
ATTORNEY           -6.1102     0.9603  -6.363 2.92e-10
SEATBELT           19.2165     4.1074   4.678 3.25e-06
ATTORNEY:SEATBELT -18.6293     8.1635  -2.282   0.0227
---
Residual standard error: 15.7 on 1087 degrees of freedom
Multiple R-squared:  0.06056,Adjusted R-squared:  0.05796
F-statistic: 23.36 on 3 and 1087 DF,  p-value: 1.182e-14
```

What is the fitted slope of the model that Austin presents to Caitlyn?

Solution: The slope is 0. This model only fits 4 difference intercept terms.

(f) (3 points) Austin wants to compare the first model that he fit and the model that he presented to Caitlyn to see whether the larger model is statistically significantly better at explaining the LOSS.

i. (1 point) What null hypothesis would Austin be testing?

Solution: $H_0 : (\beta_{\texttt{TIME}}, \beta_{\texttt{TIME:ATTORNEY}}, \beta_{\texttt{TIME:SEATBELT}}) = (0, 0, 0)$

ii. (1 point) What test statistic would Austin be using? (Specify the general form of the test statistic and identify each element. There is no need to compute the observed test statistic value.)

Solution: Let Model A be the first model and Model B be the second model. The test statistic is

$$F = \frac{(n-p-1) \cdot (RSS_B - RSS_A)}{(p-q) \cdot RSS_A},$$

where $n = 1091, p = 6, q = 3$, hence, $n - p - 1 = 1084$ and $p - q = 3$.

iii. (1 point) Under the null hypothesis, what is the distribution of Austin's test statistic? (Specify the degrees of freedom.)

Solution:) $F_{3,1084}$

(g) (Bonus: 1 points) Caitlyn is confused by the negative coefficient for the `ATTORNEY` and `SEATBELT` interaction term. Provide an interpretation to Caitlyn of the interaction term's coefficient.

Solution: There are essentialy 4 groups here: attorney used and seatbelt worn, attorney used and seatbelt not worn, attorney not used and seatbelt worn, and attorney not used and seatbelt not worn. The interaction term only comes into effect for the final group. Based on the magnitude of the interaction term relative to the seatbelt term, it appears that not having an attorney completely cancels out not wearing a seatbelt. We suspect that lawyers get involved in the claims process for severe accidents; that is, the attorney variable could be signaling a major accident versus a minor accident.

3. (5 points) (a) (1 point) Suppose you are performing 50 independent hypothesis tests each at the level $\alpha = 0.05$. What is the probability of falsely rejecting at least one of those hypothesis?

Solution: This is equal to 1 - the probability of not making any false rejections in all 50 tests: $1-(1-\alpha)^{50} \approx 0.923$.

(b) (2 points) Suppose that you would reject 50 hypotheses while controlling the FDR at $q = .1$. What is the expected number of false rejections that you would make? Also, what is the expected number of true rejections?

Solution: Controlling the FDR at 0.1 means that the expected amount of false rejections is 10%. Hence, if we reject 50 hypotheses, that means we the expected number of false rejections is $50 * 0.1 = 5$. Then the expected number of true rejections is $50 - 5 = 45$

(c) (1 point) Four independent hypothesis tests were performed and the following p-values were obtained:

- For null hypothesis A - 0.009
- For null hypothesis B - 0.027
- For null hypothesis C - 0.027
- For null hypothesis D - 0.016

Suppose you apply the Holm correction procedure with FWER control at 5%. Which null hypotheses would be rejected in this case?

Solution: We first sort the p-values:

$$0.009 < 0.016 < 0.027 < 0.027.$$

We form the appropriate adjusted significance level $0.05/(5-i), i \in \{1, \dots, 4\}$:

$$0.0125 < 0.0167 < 0.025 < 0.05.$$

Then we find the smallest p-value that is larger than its corresponding adjusted significance level. In this case that is p-value: 0.027. So we can only reject the null hypotheses corresponding to p-values 0.009 and 0.016, that is D and A. (1 P.)

(d) (1 point) Suppose you apply Benjamini-Hochberg correction procedure with FDR control at 1%, to the same hypothesis tests as above. Which null hypotheses would be rejected in this case?

Solution: We first sort the p-values:

$$0.009 < 0.016 < 0.027 < 0.027.$$

We form the appropriate adjusted significance level $0.0025 * i, i \in \{1, \dots, 4\}$:

$$0.0025 < 0.005 < 0.0075 < 0.01.$$

Then we find the largest p-value that is smaller than its corresponding adjusted significance level. In this case no p-values is smaller than its corresponding reference level. Hence, we do not reject any of the null hypothesis above. (1 P.)

4. (4 points) (a) (1 point) Which of the following statements is **true**?

1) The partial F-test which is used to compare two nested models is a two-sided test.

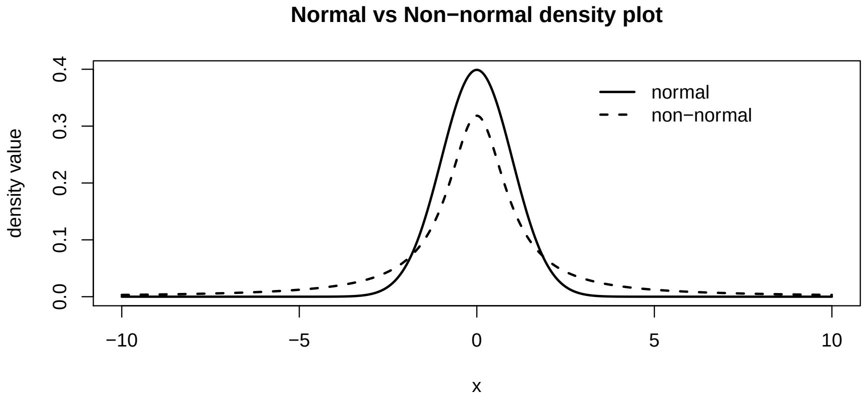
2) If the value of the test-statistics of for the global F-test is extremely small, we reject the corresponding null hypothesis.

3) Least-squares estimates $\underline{\hat{\beta}}$ in a multiple linear regression are unbiased estimators of $\underline{\beta}$.

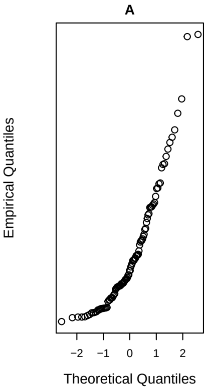
4) None of the above is true.

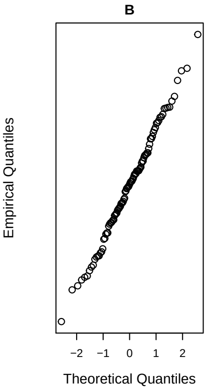
Solution: 3) See properties of OLS estimators in lectures on SLR and MLR.

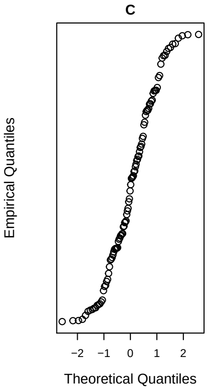
(b) (1 point) The following is a density plot comparing normal distribution vs a non-normal distribution.

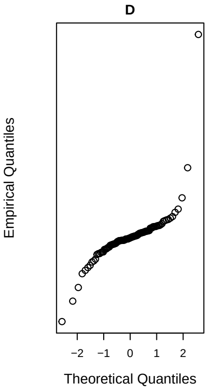


Which of the following QQ-plots given below would you expect to correspond to the non-normal distribution above?









1) A

2) B

3) C

4) D

Solution: D, because the non-normal distribution is heavy-tailed.

(c) (1 point) Assume that a simple least-squares regression model holds: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1 \dots n$ and let $\overline{x}$ and $\overline{y}$ be the empirical means for `X` and `Y` respectively. Which of the following is **true**?

1) The sum of the residuals is zero.

2) The point $(\overline{x}, \overline{y})$ is on the fitted regression line.

3) Coefficient β_1 is estimated as $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$.

4) All of the above.

Solution: 4): This is a least-squares regression, so 1), 2) and 3) are true.

(d) (1 point) Which of the following statements is **false**?

1) If a 95%-confidence interval does not contain the value 0, we can reject the null hypothesis $H_0 : \beta_j = 0$ at the 5%-level.

2) If the p-value for the hypothesis test with $H_0 : \beta_j = 0$ and $H_A : \beta_j \neq 0$ is larger than the chosen significance level, we can be sure that the corresponding variable does not have a linear influence on the response.

3) If a 95%-confidence interval does not contain the value 0, we can reject the null hypothesis $H_0 : \beta_j = 0$ at the 10%-level.

4) If the p-value for the hypothesis test with $H_0 : \beta_j = 0$ and $H_A : \beta_j \neq 0$ is smaller than the chosen significance level, the null hypothesis can be rejected.

Solution: 2): The test assesses the effect when the other predictors have been accounted for. So marginally, there can still be a linear influence on the response, even when the hypothesis test with $H_0 : \beta_j = 0$ and $H_A : \beta_j \neq 0$ is not significant.

5. (4 points) We consider a data set about housing values in the suburbs of Boston. The response variable is `MEDV` - the median value of owner-occupied homes in $1000's and there are 12 predictor variables:

`CRIM`	per capita crime rate by town
`ZN`	proportion of residential land zoned for lots over 25,000 sq.ft.
`IND`	proportion of non-retail business acres per town
`CHAS`	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
`NOX`	nitric oxides concentration (parts per 10 million)
`RM`	average number of rooms per dwelling
`AGE`	proportion of owner-occupied units built prior to 1940
`DIS`	weighted distances to five Boston employment centres
`RAD`	index of accessibility to radial highways
`TAX`	full-value property-tax rate per $10,000
`PTRATIO`	pupil-teacher ratio by town
`LSTAT`	Percentage lower status of the population

We perform a multiple linear regression and show the following shortened output, which will be used in the first three subproblems:

```
              Estimate Std. Error  Pr(>|t|)
(Intercept)  41.617270   4.936039 3.794e-16
CRIM         -0.121389   0.033000 2.605e-04
ZN            0.046963   0.013879 7.720e-04
IND           0.013468   0.062145 8.285e-01
CHAS          2.839993   0.870007 1.173e-03
NOX         -18.758022   3.851355 1.502e-06
RM            3.658119   0.420246 4.808e-17
AGE           0.003611   0.013329 7.866e-01
DIS          -1.490754   0.201623 6.171e-13
RAD           0.289405   0.066908 1.844e-05
TAX          -0.012682   0.003801 9.124e-04
PTRATIO      -0.937533   0.132206 4.630e-12
LSTAT        -0.552019   0.050659 6.392e-25

Residual standard error: 4.8 on 493 degrees of freedom
```

(a) (2 points) Which of the following statements is **true** and why?

1) The p-value for testing $\beta_{\texttt{IND}} = 0$ is significant at the 5% level.

2) The t-value associated with the intercept is approx. 4.936.

3) The two-sided 90%-confidence interval for the coefficient of `AGE` does not contain the value 0.

4) None of the above is true.

Solution: 1) is false: The p-value is $0.8285 > 0.05$.
2) is false: the t-value of the intercept is approx. 8.43.
3) is false: From the duality between the hypothesis test and the confidence interval: As the test is not significant at the 10%-level, the corresponding 90%-confidence interval does contain the zero. (Indeed, the CI is [-0.02, 0.03])

(b) (2 points) Which of the following statements is **false** and why?

1) The fitted median value of owner-occupied home (MEDV) when decreasing `NOX` by one unit, and keeping the other predictors the constant, increases by approx. \$ 18758.

2) We now remove all predictors from the above model that were not significant at the 5% level and refit the regression (no variable transformations are made). Then the F-test for a hierarchical model comparison between the two models is based on an F distribution with 2 and 493 degrees of freedom.

3) The data set contains 506 observations.

4) The regression line of the above model has an intercept of approx. 41.6173 for observations associated with `CHAS = 1`; the intercept for observations with `CHAS = 0` is approx. 44.4573.

Solution:

4) is false: The statement should be reversed: The regression line of the above model has an intercept of approx. 41.6173 for observations associated with `CHAS = 0`; the intercept for observations with `CHAS = 1` is approx. 44.4573.
2) is correct: $F = \frac{n-(p+1)}{p-q} \cdot \frac{RSS_{small}-RSS_{big}}{RSS_{big}} \sim F_{p-q,n-(p+1)}$. Here: $p-q = 12-10 = 2$ and $n-p-1 = 506-12-1 = 493$.
3) is correct as $df = n-p-1 = 493$, so n is 506.

6. (Bonus 4 points) We consider a simple linear regression with dependent variable Y and independent variable X. The estimated coefficients using the least-squares method are denoted by $\hat{\alpha}$ (intercept) and $\hat{\beta}$ (slope). Moreover, $\hat{\sigma}$ has been calculated already. We now define $X' = X - 10$ and compute the regression of Y on X'.

(a) (1 point) What is $\hat{\beta}'$ equal to (in terms of $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}$)?

Solution:

With the formula for the coefficients, we obtain

$$\begin{aligned}\hat{\beta}' &= \frac{\sum(X_i' - \bar{X}')(Y_i - \bar{Y})}{\sum(X_i' - \bar{X}')^2} = \frac{\sum((X_i - 10) - (\bar{X} - 10))(Y_i - \bar{Y})}{\sum((X_i - 10) - (\bar{X} - 10))^2} \\ &= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \hat{\beta}\end{aligned}$$

(b) (1 point) What is $\hat{\alpha}'$ equal to (in terms of $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}$)?

Solution:

With the formula for the coefficients, we obtain

$$\hat{\alpha}' = \bar{Y} - \hat{\beta}'\bar{X}' = \bar{Y} - \hat{\beta}\bar{X}' = \bar{Y} - \hat{\beta}(\bar{X} - 10) = (\bar{Y} - \hat{\beta}\bar{X}) + 10\hat{\beta} = \hat{\alpha} + 10\hat{\beta}$$

(c) (2 points) What is $\hat{\sigma}'$ equal to (in terms of $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}$)?

Solution:

With the formula for the coefficients, we obtain

$$\hat{Y}_i' = \hat{\alpha}' + \hat{\beta}'X_i' = (\hat{\alpha} + 10\hat{\beta}) + \hat{\beta}X_i' = (\hat{\alpha} + 10\hat{\beta}) + \hat{\beta}(X_i - 10) = \hat{\alpha} + \hat{\beta}X_i = \hat{Y}_i.$$

This implies that the residuals stay unchanged:

$$e_i' = Y_i - \hat{Y}_i' = Y_i - \hat{Y}_i = e_i.$$

In particular, it holds that $RSS' = \sum e_i'^2 = \sum e_i^2 = RSS$ and

$$\hat{\sigma}' = \sqrt{RSS'/(n-2)} = \sqrt{RSS/(n-2)} = \hat{\sigma}.$$