Midterm Solutions

PSTAT 120C

Summer 2022 Session B

Instructions: This exam is open book and open note and has no strict time limit. You can use any course materials; you are also allowed to use results in the book, lecture notes, and past homework without repeating proofs or derivations. Please do not consult with other students until after the submission deadline has passed.

You may use statistical software, including (but not limited to) R or Python, to help answer these questions, or you may solve them manually. If you do use software, you *must* also submit your code.

By submitting your work, you are acknowledging that your work is entirely your own.

Background

The Environmental Protection Agency, or EPA, regularly publishes data on automotive trends by year; it has maintained its database since 1975 and is updated annually to include the most up-to-date data available for all model years.

Real-world miles per gallon (*mpg*) refers to an EPA-calculated weighted average of city and highway miles per gallon. Engine displacement (*displacement*) is measured in cubic centimeters (cm^3); it is considered an expression of engine size, or a representation of the power an engine is capable of exerting and the amount of fuel it can be expected to consume.

For this exam, you'll investigate the relationship(s) between *mpg*, weight in pounds (*weight*), and *displacement*. You'll use a random sample of 15 vehicles from the summary automotive trends data, which includes information about vehicle attributes for model years ranging from 1975 to 2021.

Below, Figure 1 shows the distribution of *mpg* in a random sample of 15 vehicles.

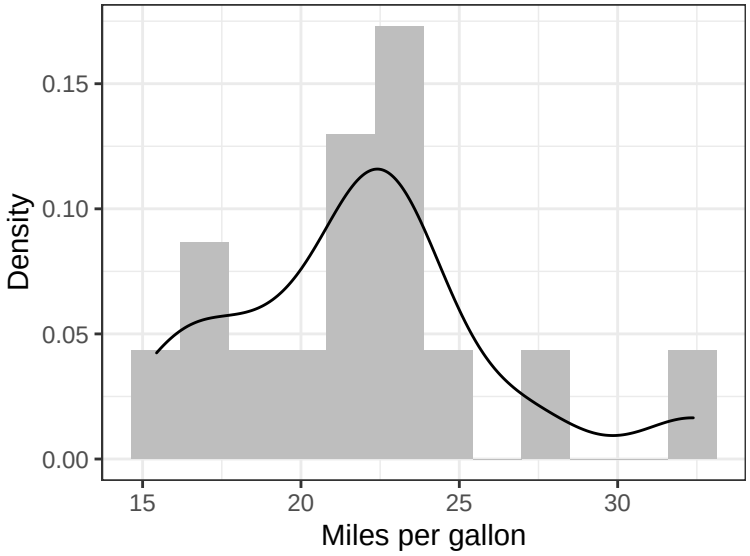


Figure 1: Histogram and density curve for miles per gallon of 15 randomly sampled vehicles.

Figure 2 is a matrix of the correlation coefficients (r^2) between *mpg*, *weight*, and *displacement*. Note that the diagonal is made up of the correlations between each variable and itself, or 1.

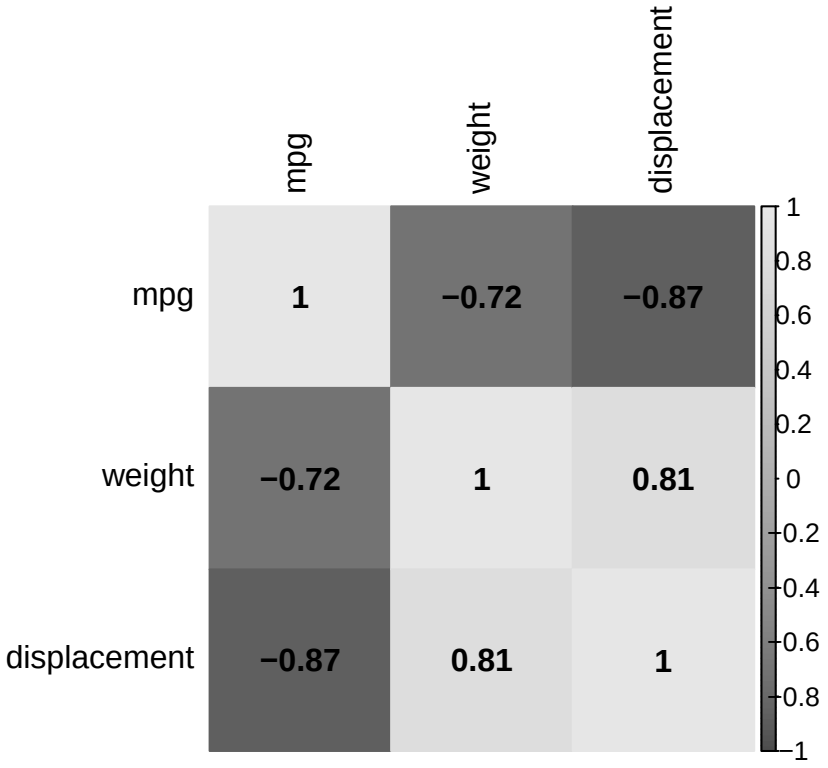


Figure 2: Correlation plot of miles per gallon, weight in pounds, and engine displacement.

Finally, the data themselves are presented in Table 1 at the end of this document. They are also available as a .csv file for download on GauchoSpace.

The overall question of interest throughout this exam is: **Should miles per gallon be predicted based on weight alone, or on the linear combination of weight and displacement?**

1. Answer the following based on a *simple* linear regression, predicting *mpg* (y) with *weight* (x_1).

   a. Fit the specified model. Write the model equation, including your estimates.

   5 pts for model fitting; 5 pts for equation

```r
data <- read_csv("cars_data.csv") %>%
select(-c(manufacturer, horsepower))

data %>%
lm(mpg ~ weight, data = .) %>%
summary()
```

```
##
## Call:
## lm(formula = mpg ~ weight, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4600 -2.1210 -0.6158  1.6716  7.0659
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.267655   5.038457   7.992 2.26e-06 ***
## weight      -0.004678   0.001267  -3.692  0.00271 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.131 on 13 degrees of freedom
## Multiple R-squared:  0.5119, Adjusted R-squared:  0.4744
## F-statistic: 13.63 on 1 and 13 DF,  p-value: 0.002709
```

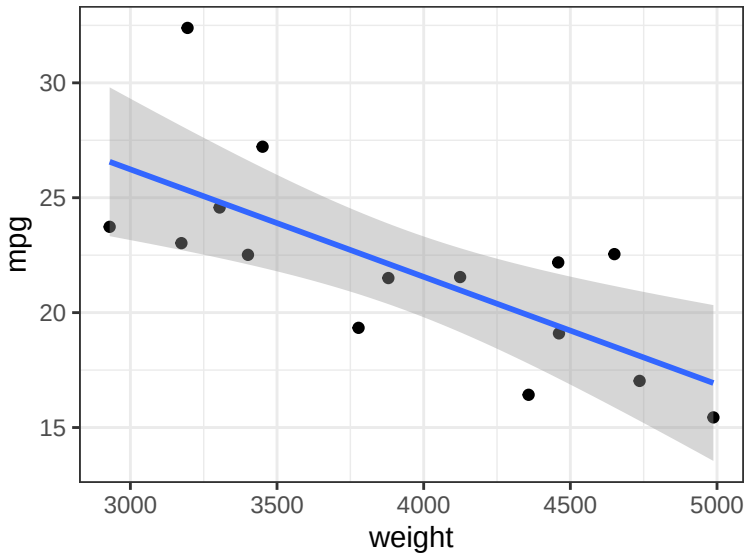
The model equation can be written as:

$$
\hat{y}_{mpg}=40.27-0.005x_{weight}
$$

   b. Create a scatterplot of *mpg* and *weight*. Add a line representing the model, with 95% confidence bands. Does the model appear to fit the data?

   5 pts for correct plot; 5 pts for discussion

```r
data %>%
ggplot(aes(x = weight, y = mpg)) +
geom_point() +
geom_smooth(method = "lm", se = T) +
theme_bw()
```



The model appears to fit the data reasonably well, although, since the sample size is small, this is difficult to assess. There are some points that appear to fall far from the line, but this is likely to be due to sampling error; overall, the model looks fine.

c. Test the null hypothesis that the slope of x_1, β_1, is equal to zero. State the hypotheses, test statistic, rejection region(s), and p-value. **Do not** interpret the conclusion of this test.

2 pts for hypotheses; 2 pts for RR; 2 pts for test statistic; 2 pts for p-value; 2 pts for not interpreting

Note that this was tested automatically when `lm()` was run; referring to those results is fine. You can also test the hypothesis manually. Either way, though, you should state everything as listed. The hypotheses and rejection region(s) are not printed by `lm()`.

The hypotheses are $H_0 : \beta_1 = 0$; $H_a : \beta_1 \neq 0$.

This is a two-tailed t-test with 13 degrees of freedom $(n - 2)$, so the rejection region is $-t_{\frac{\alpha}{2}} \cup t_{\frac{\alpha}{2}}$, or $t \leq -2.16 \cup t \geq 2.16$.

Code to fit the model, without using `lm()`:

```
X <- matrix(c(rep(1, 15), data$weight), ncol = 2)
Xt <- t(X)
SSmat <- Xt %*% X
Y <- matrix(c(data$mpg))
XYmat <- Xt %*% Y
SSmatinv <- solve(SSmat)
beta_hat <- SSmatinv %*% XYmat
beta_hat
```

```
##               [,1]
## [1,] 40.267655492
## [2,] -0.004677598
```

Note that the predicted values for β_0 and β_1 match those obtained with `lm()`. Code to conduct the hypothesis test:

```
a_mat <- matrix(c(0, 1))
at_mat <- t(a_mat)
c_val <- at_mat %*% SSmatinv %*% a_mat
yt_mat <- t(Y)
```

```r
SSE_r <- yt_mat %*% Y - (t(beta_hat) %*% XYmat)
s2 <- SSE_r/13
s <- sqrt(s2)
at_beta <- at_mat %*% beta_hat

t_val <- (at_beta - 0)/(s * sqrt(c_val)); t_val
```

```
##           [,1]
## [1,] -3.692481
```

```r
pt(q = t_val, df = 13) * 2
```

```
##              [,1]
## [1,] 0.002708602
```

The test statistic for this null hypothesis is $t = -3.69$, with a p-value of < 0.01.

2. Answer the following based on a *multiple* linear regression, predicting *mpg* with *weight* (x_1) and *engine displacement* (x_2).

   a. Fit the specified model. Write the model equation, including your estimates.

   5 pts for model fitting; 5 pts for equation

```
data %>%
lm(mpg ~ weight + displacement, data = .) %>%
summary()
```

```
##
## Call:
## lm(formula = mpg ~ weight + displacement, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1342 -0.9828 -0.6934  1.4039  5.0779
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.5095516  3.8852963   9.397 6.98e-07 ***
## weight       -0.0003083  0.0015820  -0.195    0.849
## displacement -0.0717513  0.0209294  -3.428    0.005 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.316 on 12 degrees of freedom
## Multiple R-squared:  0.7534, Adjusted R-squared:  0.7123
## F-statistic: 18.33 on 2 and 12 DF,  p-value: 0.0002248
```

   The model equation can be written as:

$$\hat{y}_{mpg} = 40.27 - 0.0003x_{weight} - 0.072x_{displ}$$

   b. Test the null hypothesis that the slope of x_1, β_1, is equal to zero. State the hypotheses, test statistic, rejection region(s), and p-value. Interpret the conclusion of this test at $\alpha = 0.05$.

   2 pts for hypotheses; 2 pts for RR; 2 pts for test statistic; 2 pts for p-value; 2 pts for not interpreting

   Note that this was tested automatically when `lm()` was run; referring to those results is fine. You can also test the hypothesis manually. Either way, though, you should state everything as listed. The hypotheses and rejection region(s) are not printed by `lm()`.

   The hypotheses are $H_0 : \beta_1 = 0$; $H_a : \beta_1 \neq 0$.

   This is a two-tailed t-test with 12 degrees of freedom $(n - 3)$, so the rejection region is $-t_{\frac{\alpha}{2}} \cup t_{\frac{\alpha}{2}}$, or $t \leq -2.18 \cup t \geq 2.18$.

   Code to fit the model, without using `lm()`:

```
X <- matrix(c(rep(1, 15), data$weight,
data$displacement), ncol = 3)
Xt <- t(X)
SSmat <- Xt %*% X
Y <- matrix(c(data$mpg))
XYmat <- Xt %*% Y
SSmatinv <- solve(SSmat)
```

```
beta_hat <- SSmatinv %*% XYmat
beta_hat
```

```
##                [,1]
## [1,] 36.5095516042
## [2,] -0.0003082554
## [3,] -0.0717512750
```

Note that the predicted values for β_0 and β_1 match those obtained with `lm()`. Code to conduct the hypothesis test:

```
a_mat <- matrix(c(0, 1, 0))
at_mat <- t(a_mat)
c_val <- at_mat %*% SSmatinv %*% a_mat
yt_mat <- t(Y)
SSE_c <- yt_mat %*% Y - (t(beta_hat) %*% XYmat)
s2 <- SSE_c/12
s <- sqrt(s2)
at_beta <- at_mat %*% beta_hat

t_val <- (at_beta - 0)/(s * sqrt(c_val)); t_val
```

```
##             [,1]
## [1,] -0.1948539
```

```
pt(q = t_val, df = 12) * 2
```

```
##           [,1]
## [1,] 0.8487674
```

The test statistic for this null hypothesis is $t = -0.195$, with a p-value of 0.85.

Interpretation: We fail to reject the null hypothesis; there is not sufficient evidence to conclude that β_1 differs from zero. In other words, there appears to be no relationship between automobile weight and miles per gallon *after controlling for engine displacement.*

c. Consider $x_1^* = 3000$ and $x_2^* = 150$. Calculate a 95% confidence interval for $E[Y|x_1 = x_1^*, x_2 = x_2^*]$. Calculate a 95% prediction interval for y_i, given $x_1 = x_1^*$ and $x_2 = x_2^*$. Interpret both of these intervals in context.

2.5 pts each for CI and PI; 2.5 pts each for interpretation

These results can be obtained in multiple ways. The solutions present two methods of doing so for each interval.

For the **confidence interval**:

```
a_xstar_mat <- matrix(c(1, 3000, 150)); a_xstar_mat
```

```
##      [,1]
## [1,]    1
## [2,] 3000
## [3,]  150
```

```
at_xstar_mat <- t(a_xstar_mat)
at_xstar_beta <- at_xstar_mat %*% beta_hat
c_val <- at_xstar_mat %*% SSmatinv %*% a_xstar_mat

ci_term <- qt(0.025, df = 12, lower.tail = F) * s * sqrt(c_val)
ci_low <- at_xstar_beta - ci_term
```

```
ci_high <- at_xstar_beta + ci_term
ci_low; ci_high
```

```
##          [,1]
## [1,] 22.35674
```

```
##          [,1]
## [1,] 27.28745
```

It's also possible to use the `predict()` function, with notably fewer lines of code, to obtain the confidence interval:

```
pred_dat <- tibble(weight = 3000, displacement = 150)
predict(lm(mpg ~ weight + displacement, data = data),
newdata = pred_dat, interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 24.82209 22.35674 27.28745
```

Regarding interpretation, there are multiple valid answers. One example is: "We can be 95% confident that the interval $(22.36, 27.87)$ contains the population mean of miles per gallon for cars that weigh $\$3,000$ pounds with an engine displacement of 150 cubic centimeters."

An answer which treats the parameter, $\mu_{Y|x=x^*}$, as a random variable is not correct; answers should recognize that the upper and lower bounds of the confidence interval themselves are random variables, not the parameter of interest.

For the **prediction interval**:

```
c_val <- 1 + at_xstar_mat %*% SSmatinv %*% a_xstar_mat

pi_term <- qt(0.025, df = 12, lower.tail = F) * s * sqrt(c_val)
pi_low <- at_xstar_beta - pi_term
pi_high <- at_xstar_beta + pi_term
pi_low; pi_high
```

```
##          [,1]
## [1,] 19.20524
```

```
##          [,1]
## [1,] 30.43895
```

It's also possible to use the `predict()` function, with notably fewer lines of code, to obtain the prediction interval:

```
predict(lm(mpg ~ weight + displacement, data = data),
newdata = pred_dat, interval = "prediction", level = 0.95)
```

```
##        fit      lwr      upr
## 1 24.82209 19.20524 30.43895
```

There are again multiple valid statements regarding interpretation. An example: "We can expect that 95% of future cars weighing $\$3,000$ pounds with an engine displacement of 150 cubic centimeters will have between 19.21 and 30.44 miles per gallon."

d. Which model constitutes the "complete" model and which the "reduced" model? Can x_2 be dropped from the model without losing predictive information? Test at the $\alpha = 0.05$ significance level.

5 pts for correct test & conclusion

The multiple regression including both x_1 and x_2 should be regarded as the "complete" model and the simple linear regression with only x_1 the "reduced."

Although not specifically asked to do so, answers should ideally include the hypotheses, rejection region(s), test statistic, and p-value.

$H_0 : \beta_2 = 0$; $H_A : \beta_2 \neq 0$.

The rejection region is $F > F_\alpha$. Since the number of independent variables in the complete model is $k = 2$ and the number in the reduced model is $g = 1$, the degrees of freedom are $k - g = 1$ for the numerator and $n - (k + 1) = 15 - 3 = 12$ for the denominator. Therefore, the rejection region is $F > 4.75$.

Notice that the SSE values for each model were stored as `SSE_r` and `SSE_c` earlier. Then the test statistic can be calculated:

```
f_stat <- (SSE_r - SSE_c)/( (SSE_c)/12 )
f_stat
```

```
##          [,1]
## [1,] 11.75288
```

And its p-value:

```
pf(f_stat, df1 = 1, df2 = 12, lower.tail = F)
```

```
##             [,1]
## [1,] 0.005001817
```

We can therefore reject the null hypothesis; there is sufficient evidence to conclude that including engine displacement results in a significantly better fit of the model to the data ($F = 11.75, p < .01$).

This test can also be conducted with `anova()`:

```
complete_m <- data %>%
lm(mpg ~ weight + displacement, data = .)

reduced_m <- data %>%
lm(mpg ~ weight, data = .)

anova(complete_m, reduced_m)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ weight + displacement
## Model 2: mpg ~ weight
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     12  64.386
## 2     13 127.445 -1    -63.06 11.753 0.005002 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3. Consider your answers to the previous questions, then answer the following.

   Suppose that the true population relationship is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

   Further suppose that there is a relationship between x_1 and x_2, given by:

$$x_2 = \gamma_0 + \gamma_1 x_1 + \delta$$

   where γ_1 and β_2 are non-zero.

   a. Find the expected values of $\hat{\beta}_0$ and $\hat{\beta}_1$ if the independent variable x_2 is omitted from the regression.

   5 pts for OLS equation for β_1; 5 pts for $E[\beta_1]$; 5 pts for $E[\beta_0]$

   It is crucial to note that the **population model** is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ while the **regression being fit** is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1$.

   Let $x = x_1$ and $z = x_2$ for simplicity.

   The OLS estimator for $\hat{\beta}_1$ is:

$$\hat{\beta}_1 = \frac{Cov(x_i, y_i)}{Var(x_i)} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{1}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x}) x_i} \tag{2}$$

   We can now substitute in the **true** definition of y_i in the **population**:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i)}{\sum_{i=1}^{n} (x_i - \bar{x}) x_i} \tag{3}$$

$$= \frac{\beta_0 \sum_{i=1}^{n} (x_i - \bar{x}) + \beta_1 \sum_{i=1}^{n} (x_i - \bar{x}) x_i + \beta_2 \sum_{i=1}^{n} (x_i - \bar{x}) z_i + \sum_{i=1}^{n} (x_i - \bar{x}) \epsilon_i}{\sum_{i=1}^{n} (x_i - \bar{x}) x_i} \tag{4}$$

$$\hat{\beta}_1 = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} (x_i - \bar{x}) z_i}{\sum_{i=1}^{n} (x_i - \bar{x}) x_i} + \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \epsilon_i}{\sum_{i=1}^{n} (x_i - \bar{x}) x_i} \tag{5}$$

   This is the OLS equation for $\hat{\beta}_1$ if we **do not** include $z = x_2$ in the model. Note that:

$$\hat{\beta}_1 = \beta_1 + \beta_2 \gamma_1 + \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \epsilon_i}{\sum_{i=1}^{n} (x_i - \bar{x}) x_i}$$

Its expected value is then:

$$E\left[\hat{\beta}_1\right] = E\left[\beta_1 + \beta_2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})\, z_i}{\sum_{i=1}^{n}(x_i - \bar{x})\, x_i} + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\, \epsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})\, x_i}\right] \tag{6}$$

$$= E\left[\beta_1\right] + E\left[\beta_2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})\, z_i}{\sum_{i=1}^{n}(x_i - \bar{x})\, x_i}\right] + E\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})\, \epsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})\, x_i}\right] \tag{7}$$

$$E\left[\hat{\beta}_1\right] = \beta_1 + \beta_2\gamma_1 \tag{8}$$

Because $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$, we see that:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{9}$$

$$\bar{y} = \beta_0 + \beta_1\bar{x} + \beta_2\bar{z} + \bar{\epsilon} \tag{10}$$

And $z_i = \gamma_0 + \gamma_1 x_i + \delta_i$, so:

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i \tag{11}$$

$$\bar{z} = \gamma_0 + \gamma_1\bar{x} + \bar{\delta} \tag{12}$$

Then the expected value of $\hat{\beta}_0$ is:

$$E\left[\hat{\beta}_0\right] = E\left[\bar{y}\right] - E\left[\hat{\beta}_1\bar{x}\right] \tag{13}$$

$$= E\left[\beta_0 + \beta_1\bar{x} + \beta_2\left(\gamma_0 + \gamma_1\bar{x} + \bar{\delta}\right) + \bar{\epsilon}\right] - E\left[\hat{\beta}_1\bar{x}\right] \tag{14}$$

$$= \beta_0 + \beta_1\bar{x} + \beta_2\gamma_0 + \beta_2\gamma_1\bar{x} - \bar{x}E\left[\hat{\beta}_1\right] \tag{15}$$

$$= \beta_0 + \beta_1\bar{x} + \beta_2\gamma_0 + \beta_2\gamma_1\bar{x} - \bar{x}\left(\beta_1 + \beta_2\gamma_1\right) \tag{16}$$

$$= \beta_0 + \beta_1\bar{x} + \beta_2\gamma_0 + \beta_2\gamma_1\bar{x} - \bar{x}\beta_1 - \bar{x}\beta_2\gamma_1 \tag{17}$$

$$E\left[\hat{\beta}_0\right] = \beta_0 + \beta_2\gamma_0 \tag{18}$$

b. Calculate the bias (if any) of $\hat{\beta}_0$ and $\hat{\beta}_1$ when x_2 is omitted.

5 pts each for bias equations

For $\hat{\beta}_0$:

$$Bias\left(\hat{\beta}_0\right) = E\left[\hat{\beta}_0\right] - \beta_0 \tag{19}$$

$$= \beta_0 + \beta_2\gamma_0 - \beta_0 \tag{20}$$

$$Bias\left(\hat{\beta}_0\right) = \beta_2\gamma_0 \tag{21}$$

And for $\hat{\beta}_1$:

$$Bias\left(\hat{\beta}_1\right) = E\left[\hat{\beta}_1\right] - \beta_1 \tag{22}$$

$$= \beta_1 + \beta_2\gamma_1 - \beta_1 \tag{23}$$

$$Bias\left(\hat{\beta}_1\right) = \beta_2\gamma_1 \tag{24}$$

c. What values of γ_1 and β_2 would result in $\hat{\beta}_0$ and $\hat{\beta}_1$ remaining unbiased?

2 pts for valid answer

For $\hat{\beta}_0$ and $\hat{\beta}_1$ to remain unbiased, β_2 must equal zero or both γ_0 and γ_1 must equal zero.

In light of the above:

i. What assumption of linear regression is being violated in Question 1? Is this assumption met in Question 2?

4 pts

By fitting a regression including only x_1, we are violating the assumption that the error terms, ϵ_i, are uncorrelated with the explanatory variable(s). In other words, the assumption that $E\left(\epsilon_i|x_1\right) = 0$ is violated in Question 1.

Arguably, this assumption is met in Question 2, because we included x_2. (However, there may be other variables correlated with miles per gallon, weight, and engine displacement which are not included in Question 2; in other words, it's technically possible that Question 2 suffers from omitted variable bias as well.)

ii. In Question 1, are the estimates of β_0 and β_1 BLUE? Why or why not?

4 pts

They are not, because – as demonstrated – the estimators in Question 1 are not unbiased.

Table 1. Raw data for a random sample of 15 vehicles from the EPA Automative Trends Database.

Model Year	MPG	Weight	Displacement	Class
2015	21.54716	4124.129	178.5575	Truck
2007	17.02911	4736.041	236.0139	Truck
2003	19.33781	3777.898	179.4107	Truck
1986	23.02399	3174.024	190.2972	Car
2017	22.54566	4650.112	164.4554	Truck
Prelim. 2021	32.38923	3194.868	114.4701	Car
2000	22.51440	3400.909	168.2990	Car
2014	22.18444	4458.683	208.4433	Truck
1999	21.50476	3879.585	197.3525	Car
2012	27.21958	3450.740	137.7964	Car
1990	23.73371	2929.358	122.0215	Car
1998	24.57349	3304.248	142.4937	Car
2007	19.09633	4461.215	218.8619	Truck
2002	15.44052	4987.675	302.1571	Truck
1988	16.42429	4357.654	239.6896	Truck