Homework-4-Q1

BIOS507 Spring 2025 | Dr Lukemire | Elizabeth Nemeti Due: March 17 2025

Problem 1.

Researchers collected a random sample of data on infants' birth weights (Y , lbs), gestation period (X1, weeks), and a variable whose value is the number of the letter of the alphabet the baby's last name starts with (A =1, B=2, C=3, etc). Treat X2 as a quantitative variable. The data set is birth_weight.txt.

- **Y**: infants' birth weights in lbs
- **X1**: gestation period in weeks
- **X2**: quantitative variable whose value is the number of the letter of the alphabet the baby's last name starts with (A=1, B=2, C=3, etc)

```
data_path = "/Users/elizabethnemeti/Documents/GitHub/BIOS507-Coursework/Homeworks/Homework-4/"
data_file <- file.path(data_path, "birth_weight.txt")
infant_data <- read.delim(data_file, header = TRUE)

head(infant_data)
```

```
##      Y X1 X2
## 1 6.75 36 11
## 2 8.00 39 14
## 3 7.50 38  4
## 4 6.50 36  2
## 5 7.25 37 23
## 6 7.00 37  1
```

```
str(infant_data)
```

```
## 'data.frame':    10 obs. of  3 variables:
##  $ Y : num  6.75 8 7.5 6.5 7.25 7 5.5 7.5 8 6.75
##  $ X1: int  36 39 38 36 37 37 35 38 39 36
##  $ X2: int  11 14 4 2 23 1 24 5 15 12
```

Now that we're doing multiple linear regression, we'll have multiple predictor variables unlike in simple linear regression where we just had one X.

predictor variable (X1) -> gestation period (weeks) predictor variable (X2) -> numeric code for baby's last name initial (A=1, B=2, etc.) response variable (Y) -> infants' birth weights in lbs

Model -> Y = ß0 + ß1(X1) + ß2(X2) + E

1. Run a regression of Y on X1 and X2. Is the improvement due to the additional of X2 (to a model already including X1 significant? Use = 0.05 for the test.

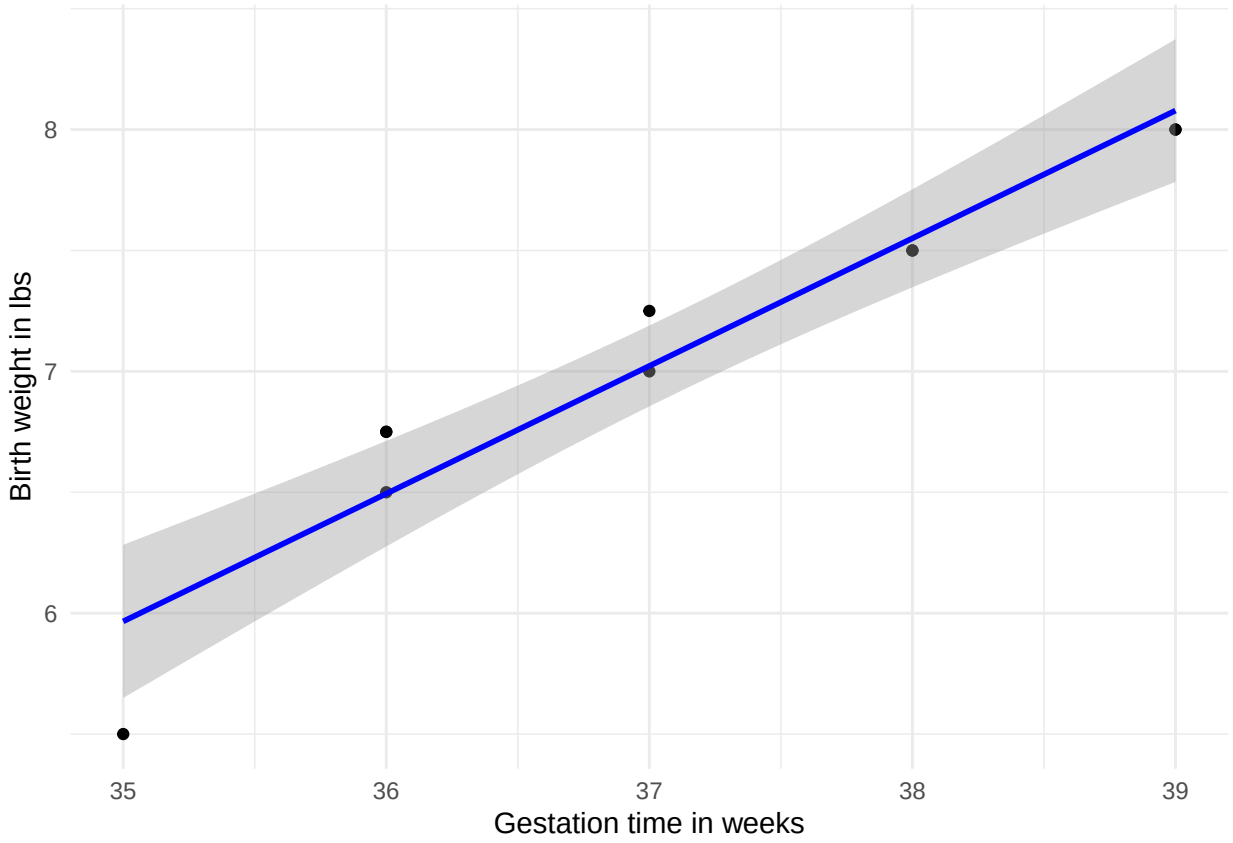
```
summary(infant_data)
```

```
##        Y               X1              X2
##  Min.   :5.500   Min.   :35.0   Min.   : 1.00
##  1st Qu.:6.750   1st Qu.:36.0   1st Qu.: 4.25
##  Median :7.125   Median :37.0   Median :11.50
##  Mean   :7.075   Mean   :37.1   Mean   :11.10
```

```
##  3rd Qu.:7.500   3rd Qu.:38.0   3rd Qu.:14.75
##  Max.   :8.000   Max.   :39.0   Max.   :24.00
```

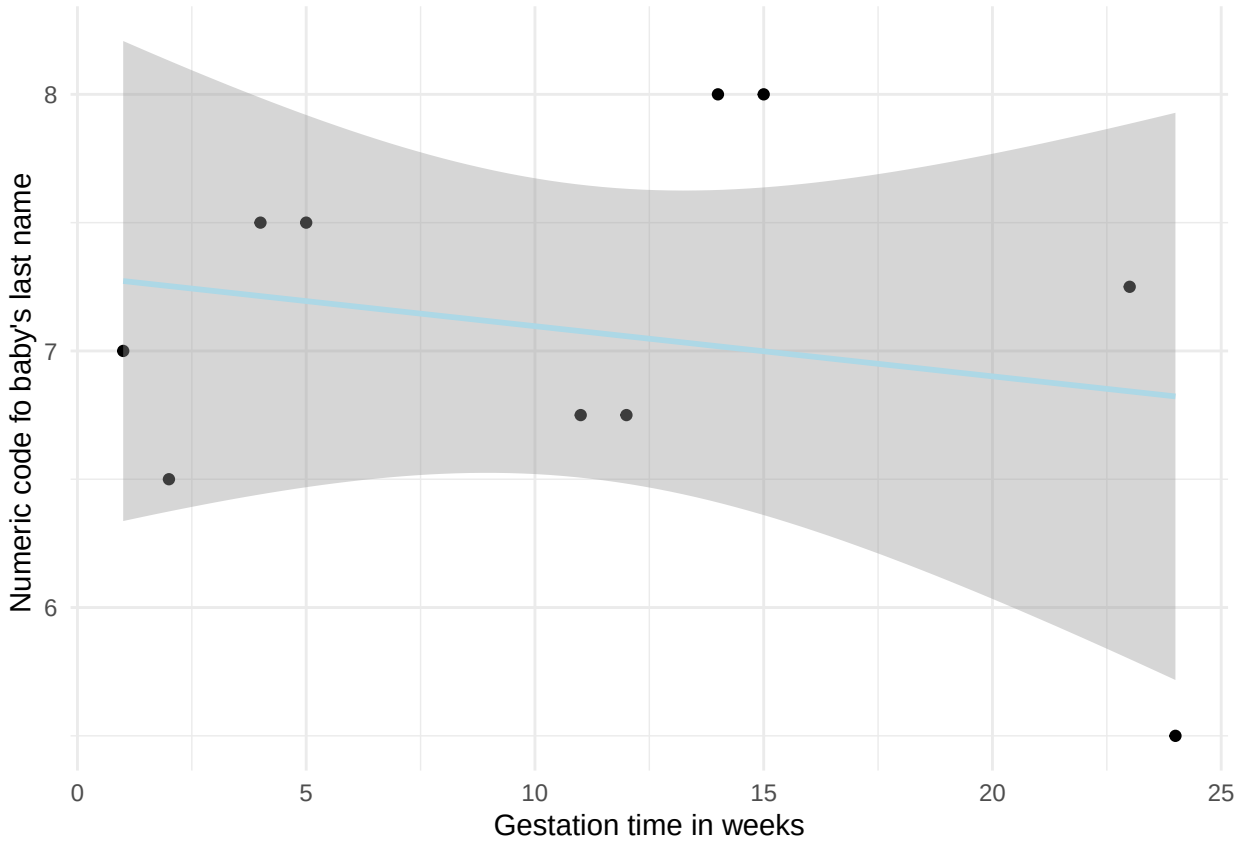
```
infant_data_scatterplot1 <- ggplot(
  infant_data,
  aes(x = X1, y = Y)) +
  geom_point() +
  geom_smooth(method = "lm", color = "blue") +
  theme_minimal() +
  xlab("Gestation time in weeks") +
  ylab("Birth weight in lbs")
infant_data_scatterplot1
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
infant_data_scatterplot2 <- ggplot(
  infant_data,
  aes(x = X2, y = Y)) +
  geom_point() +
  geom_smooth(method = "lm", color = "lightblue") +
  theme_minimal() +
  xlab("Gestation time in weeks") +
  ylab("Numeric code fo baby's last name ")
infant_data_scatterplot2
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Looking at these two plots X1, the gestation weeks, gives us a clear linear relationship, but X2, the alphabet code, does not exhibit any clear relationship and probably won't add explanatory power.

```
model_reduced <- lm(Y ~ X1, data = infant_data)
summary(model_reduced)
```

```
##
## Call:
## lm(formula = Y ~ X1, data = infant_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46598 -0.07138 -0.03624  0.17234  0.25592
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12.5177     2.0639  -6.065 0.000301 ***
## X1            0.5281     0.0556   9.499 1.24e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2286 on 8 degrees of freedom
## Multiple R-squared:  0.9186, Adjusted R-squared:  0.9084
## F-statistic: 90.23 on 1 and 8 DF,  p-value: 1.244e-05
```

```
model_full <- lm(Y ~ X1 + X2, data = infant_data)
summary(model_full)
```

```
##
```

```
## Call:
## lm(formula = Y ~ X1 + X2, data = infant_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41473 -0.07391 -0.05309  0.17695  0.28392
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12.286444   2.223984  -5.525 0.000883 ***
## X1            0.523295   0.059337   8.819 4.87e-05 ***
## X2           -0.004756   0.009918  -0.479 0.646211
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2404 on 7 degrees of freedom
## Multiple R-squared:  0.9211, Adjusted R-squared:  0.8986
## F-statistic: 40.89 on 2 and 7 DF,  p-value: 0.0001377
```

```
anova(model_reduced, model_full)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1
## Model 2: Y ~ X1 + X2
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1      8 0.41790
## 2      7 0.40461  1  0.013289 0.2299 0.6462
```

H0: ß2 = 0 (aka if we have X1 already in the model, X2 does not help predict Y) HA: ß2 != 0 (aka if we have X1 already in the model, X2 does help predict Y) = 0.05

Since p = 0.6462 > 0.05 there is no evidence that X2 improves our prediction of birth weight once X1 is accounted for, therefore, we fail to reject the null hypothesis.

2. Is the previous result surprising to you? Why or why not?

Not surprising! If we first consider why someone might collect these variables in particular, it's clear that gestation time (X1) is a meaningful predictor since it reflects a baby's growth over time. However, the alphabet/numeric code for the baby's last name (X2) has no intuitive connection to birth weight. It appears to rather be noise, which was indicated with the scatterplot for X2, which did not show a meaningful linear relationship.

3. Calculate the square of the partial correlations between Y and X1 given X2 (R2 Y,X1|X2) and between Y and X2 given X1 (R2 Y,X2|X1)

First, let's find our t statistics.

```
model_full <- lm(Y ~ X1 + X2, data = infant_data)
summary(model_full)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = infant_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41473 -0.07391 -0.05309  0.17695  0.28392
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12.286444   2.223984  -5.525 0.000883 ***
## X1            0.523295   0.059337   8.819 4.87e-05 ***
## X2           -0.004756   0.009918  -0.479 0.646211
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2404 on 7 degrees of freedom
## Multiple R-squared:  0.9211, Adjusted R-squared:  0.8986
## F-statistic: 40.89 on 2 and 7 DF,  p-value: 0.0001377
```

X1 t-value = 8.819 X2 t-value = -0.479 df = 7

formula: square root((t value of X1^2) / (t value of X1^2 + df))

```
df <- 7

# for X1, controlling for X2
t_X1 <- 8.819
r_X1_given_X2 <- sqrt((t_X1^2) / (t_X1^2 + df))
r_X1_given_X2_sq <- r_X1_given_X2^2
r_X1_given_X2_sq # the partial R^2 for X1
```

```
## [1] 0.9174283
```

```
# for X2, controlling for X1
t_X2 <- -0.479
r_X2_given_X1 <- sqrt((t_X2^2) / (t_X2^2 + df))
r_X2_given_X1_sq <- r_X2_given_X1^2
r_X2_given_X1_sq  # the partial R^2 for X2
```

```
## [1] 0.03173703
```

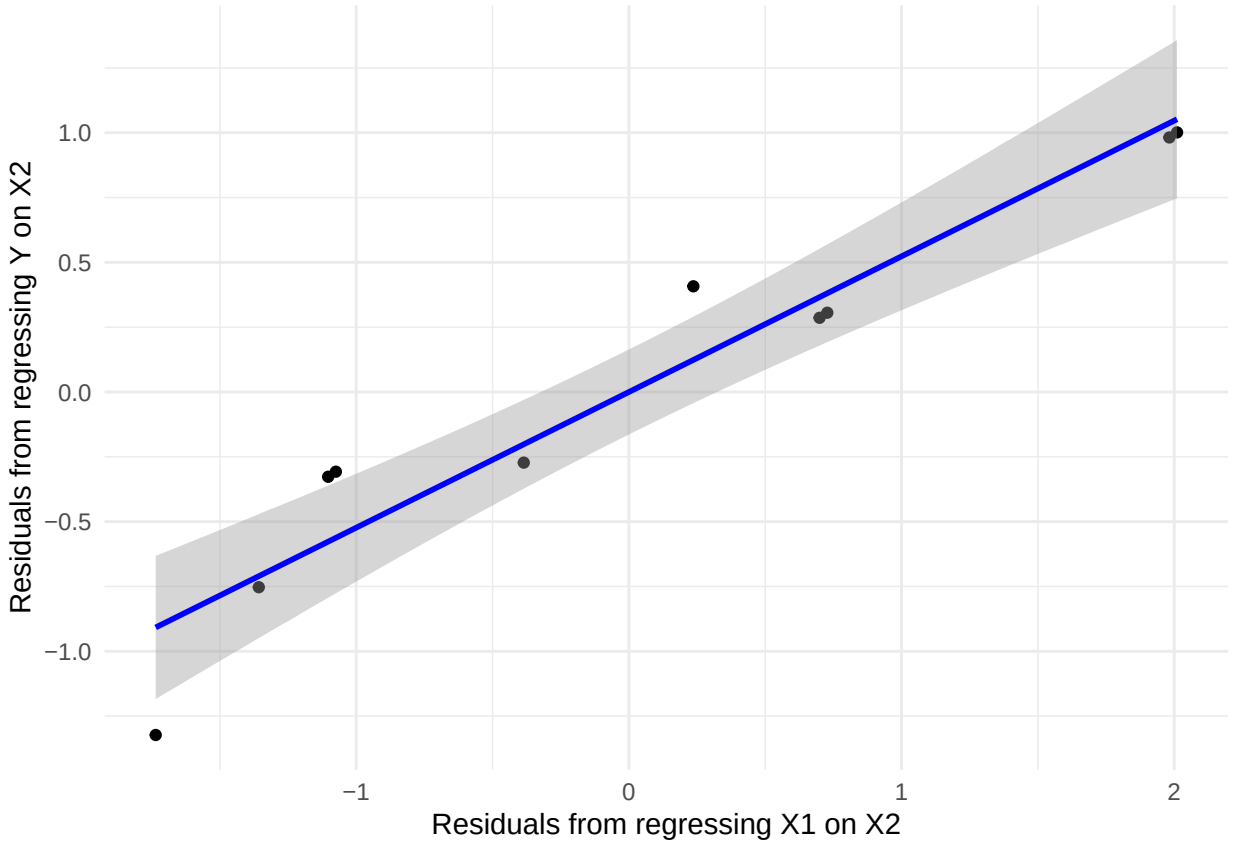
Our partial R^2 for X1 is 0.9174283, where X1 explains 91.74% of the remaining variation after X2 is controlled for. Our partial R^2 for X2 is 0.03173703, where X2 explains 3.17% of the remaining variation after X1 is controlled for. This is what we would expect considering gestation weeks (X1) hold more explanatory power over a "noise" variable (X2).

4. Run two different regression models: (a) A simple linear regression of Y onto X2 (b) A simple linear regression of X1 onto X2. Then, produce a plot of the two sets of residuals against each other (this plot will have the residuals from model (a) on the y-axis and the residuals from model (b) on the x-axis).

```
# (a) simple linear regression of Y onto X2
model_a <- lm(Y ~ X2, data = infant_data)

# (b) simple linear regression of X1 onto X2
model_b <- lm(X1 ~ X2, data = infant_data)

residuals_a <- resid(model_a) # get residuals for Y onto X2
residuals_b <- resid(model_b) # get residuals for X1 onto X2
residuals_data <- data.frame(residuals_a, residuals_b)

ggplot(residuals_data,
       aes(x = residuals_b,     # (b) on the x-axis
           y = residuals_a)) + # (a) on the y-axis
```

```
  geom_point() +
  geom_smooth(method = "lm", color = "blue") +
  theme_minimal() +
  xlab("Residuals from regressing X1 on X2") +
  ylab("Residuals from regressing Y on X2")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



As the plot shows the relationship between Y and X1 after the influence of X2 has already been controlled for, we can infer that there is still clearly a linear relationship which is expected based on our initial analyses on whether X2 was a meaningful variable or not (it wasn't), and, our partial R^2 indicating X2 holds little explanatory power.

5. Next, run a simple linear regression of the residuals from model (a) against the residuals from model (b) obtained above. What is the estimated slope and the R2 for this simple linear regression? Also, what regression coefficient and squared partial correlation coefficient from your work with the full model in part 1 do these quantities correspond to? Why does this connection make sense?

```
model_residuals <- lm(residuals_a ~ residuals_b, data = residuals_data)
summary(model_residuals)
```

```
##
## Call:
## lm(formula = residuals_a ~ residuals_b, data = residuals_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41473 -0.07391 -0.05309  0.17695  0.28392
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.231e-16  7.112e-02   0.000        1
## residuals_b 5.233e-01  5.550e-02   9.428 1.32e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2249 on 8 degrees of freedom
## Multiple R-squared:  0.9174, Adjusted R-squared:  0.9071
## F-statistic: 88.89 on 1 and 8 DF,  p-value: 1.315e-05
```

What is the estimated slope and the R2 for this simple linear regression?

slope -> 5.233e-01 R^2 -> 0.9174

What regression coefficient and squared partial correlation coefficient from your work with the full model in part 1 do these quantities correspond to?

The regression coefficient corresponds to our estimate for X1 (0.523295) in the full model. The squared partial correlation coefficient corresponds to our partial R^2 for Y and X1 given X2 (0.9174283).

Why does this connection make sense?

By removing the effect of X2 (controlling for it using the residuals), we isolated the direct link between X1 and Y. This demonstrated that the majority of the predictive power for Y comes from X1, not X2 (which is just noise). Once we removed the influence of X2, the remaining variation in Y was explained almost entirely (~91%) by X1. Therefore, the relationship we observed between Y and X1 (with X2 being controlled for), in both the residuals plot and the full model reflects the true effect of X1 on Y, which is why the estimated slope and the R^2 from the residual regression match the coefficient and squared partial correlation from the full model. This connection is important as it confirms X1 is our true predictive variable for Y here.