Garbage Can Regression Challenge

Garbage Can Regression Challenge

```
Warning: package 'ggplot2' was built under R version 4.4.3
```

```
Warning: package 'tidyr' was built under R version 4.4.3
```

```
Warning: package 'dplyr' was built under R version 4.4.3
```

```
Warning: package 'lubridate' was built under R version 4.4.3
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
# A tibble: 15 x 4
   Stress StressSurvey  Time Anxiety
    <dbl>        <dbl> <dbl>   <dbl>
 1      0            0     0     0
 2      0            0     1     0.1
 3      0            0     1     0.1
 4      1            3     1     1.1
 5      1            3     1     1.1
 6      1            3     1     1.1
 7      2           6   2      2.2
 8      2           6   2      2.2
 9      2           6   2      2.2
10      8           9   2      8.2
11      8           9   2      8.2
12      8           9   2.1    8.21
13     12          12   2.2   12.2
14     12          12   2.2   12.2
15     12          12   2.2   12.2
```

My Analysis

Question 1: Bivariate Regression Analysis with StressSurvey: Run a bivariate regression of Anxiety on StressSurvey. What are the estimated coefficients? How do they compare to the true relationship?

The regression of Anxiety on StressSurvey produced an intercept of roughly –1.5 and a slope of 1.05, with a very high R^2. This suggests StressSurvey is a strong predictor of Anxiety. However, the true data-generating process is Anxiety = Stress + 0.1 * Time, and StressSurvey is not part of that equation. The strong association occurs only because StressSurvey is highly correlated with Stress. The model appears statistically sound, but it is attributing causality to the wrong variable, making it a classic example of a misleading regression.

```
Call:
lm(formula = Anxiety ~ StressSurvey, data = observDF)

Residuals:
   Min     1Q Median     3Q    Max 
-2.558 -0.517  0.301  1.180  1.624 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -1.5240     0.7069  -2.156   0.0504 .  
StressSurvey   1.0470     0.0962  10.883 6.68e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.581 on 13 degrees of freedom
Multiple R-squared:  0.9011,    Adjusted R-squared:  0.8935 
F-statistic: 118.4 on 1 and 13 DF,  p-value: 6.681e-08
```

```
# A tibble: 2 x 5
  term         estimate std.error statistic      p.value
  <chr>           <dbl>     <dbl>     <dbl>        <dbl>
1 (Intercept)     -1.52    0.707      -2.16 0.0504
2 StressSurvey     1.05    0.0962     10.9  0.0000000668
```

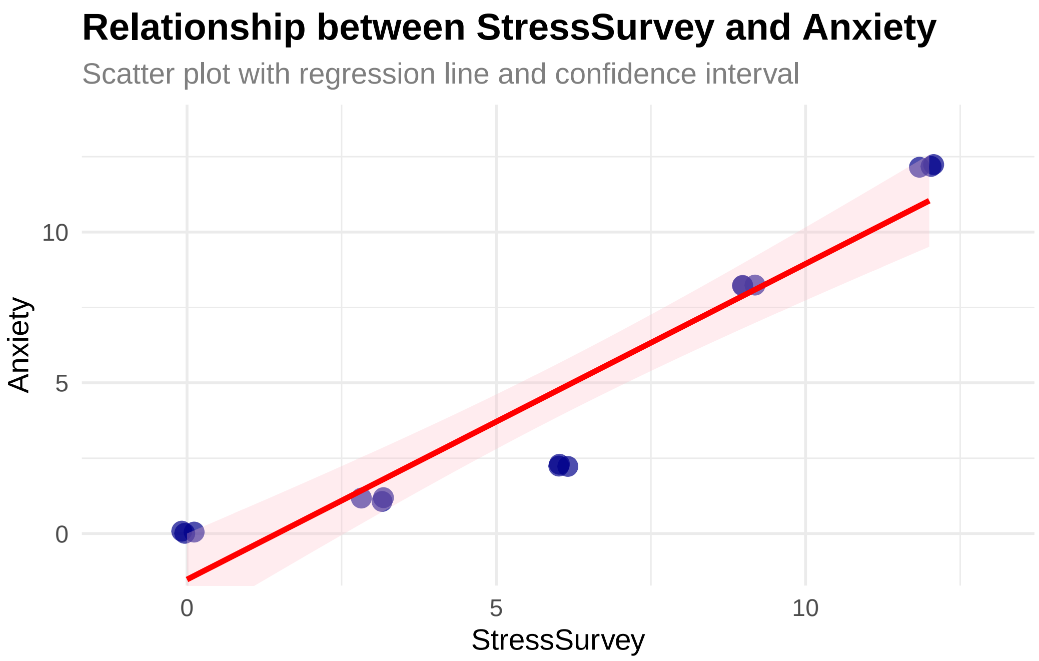
```
True relationship: Anxiety = Stress + 0.1 × Time
```

```
Estimated relationship: Anxiety = -1.524 + 1.047 × StressSurvey
```

Question 2: Visualization of Bivariate Relationship: Create a scatter plot with the regression line showing the relationship between StressSurvey and Anxiety. Comment on the fit of the model.

The scatterplot exhibits an almost perfect linear pattern, and the fitted regression line fits well with the points, further giving the impression of a strong relationship. In fact, the pattern in the data is visually quite convincing, but it involves correlation rather than causation: StressSurvey is simply proxying for Stress; this creates the illusion of a meaningful relationship when none really exists.

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



Relationship between StressSurvey and Anxiety

Scatter plot with regression line and confidence interval

Question 3: Bivariate Regression Analysis with Time: Run a bivariate regression of Anxiety on Time. What are the estimated coefficients? How do they compare to the true relationship?

The regression of Anxiety on Time produced an intercept of approximately –3.68 and a slope of 5.34, greatly exaggerating the impact of Time. In the actual model, Time only adds 0.1 to Anxiety; thus, this estimate is inflated due to omitted variable bias. Because the model does not control for Stress-the actual leading factor in Anxiety-the excess explanatory power is incorrectly given to Time.

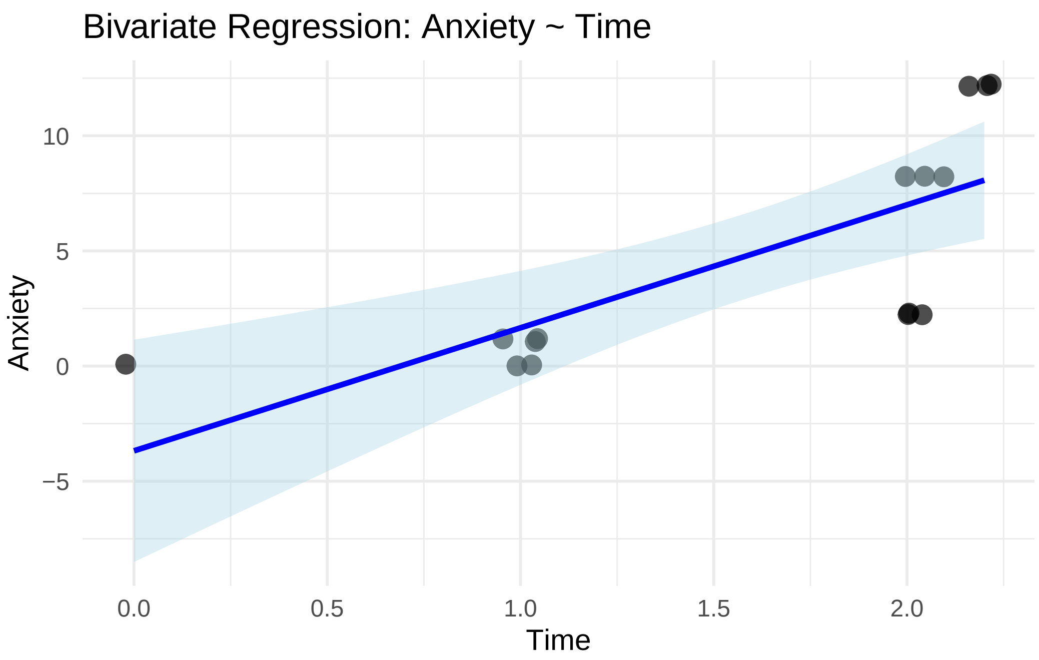
```
Call:
lm(formula = Anxiety ~ Time, data = observDF)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.8010 -1.5605 -0.5605  2.4395  4.1508 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   -3.680      2.233  -1.648  0.12330   
Time           5.341      1.305   4.093  0.00127 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.323 on 13 degrees of freedom
Multiple R-squared:  0.563, Adjusted R-squared:  0.5294 
F-statistic: 16.75 on 1 and 13 DF,  p-value: 0.00127
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)    -3.68      2.23     -1.65 0.123  
2 Time            5.34      1.30      4.09 0.00127
```

```
True relationship: Anxiety = Stress + 0.1 × Time
```

```
Estimated relationship: Anxiety = -3.6801 + 5.3406 × Time
```

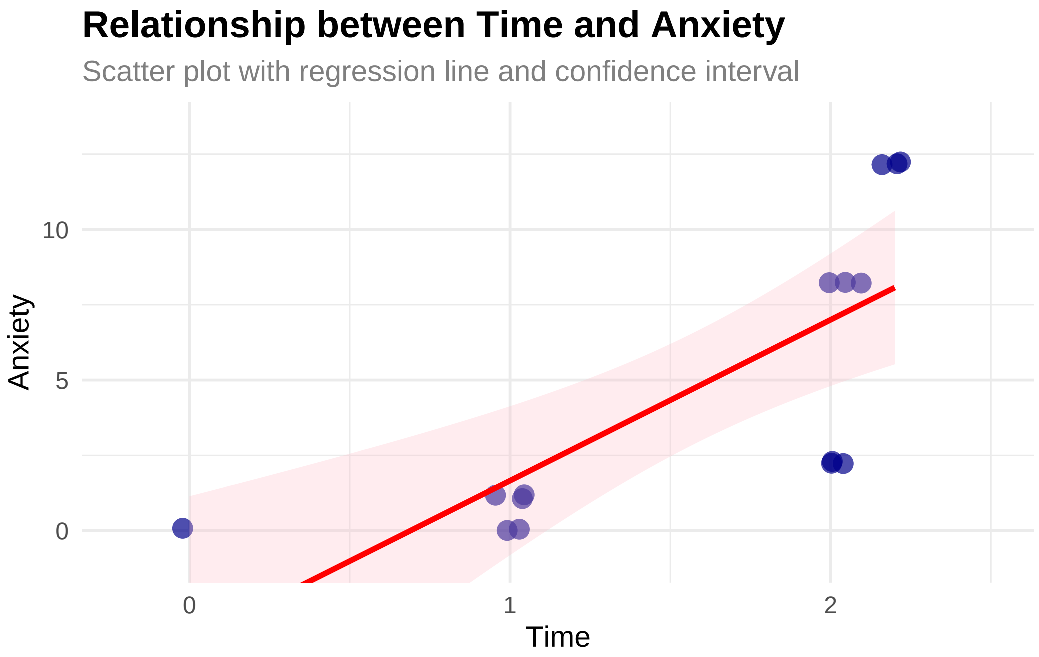
```
`geom_smooth()` using formula = 'y ~ x'
```



Question 4: Visualization of Bivariate Relationship: Create a scatter plot with the regression line showing the relationship between Time and Anxiety. Comment on the fit of the model.

The Time-Anxiety scatterplot is more diffuse, with points much less closely-packed around the regression line. While the overall tendency of the trend is positive, the weaker visual alignment reflects a much poorer R squared, indicating that Time alone supplies limited explanatory value, which emphasizes the need to incorporate Stress when modeling Anxiety.

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



Question 5: Multiple Regression Analysis: Run a multiple regression of Anxiety on StressSurvey and Time. What are the estimated coefficients? How do they compare to the true relationship?

Including both StressSurvey and Time raises R square to about 0.94, and both appear significant. But the coefficient on Time is negative, which goes against the actual relationship. This is because StressSurvey is an imperfect proxy for Stress and thus ends up causing multicollinearity and incorrect signs on coefficients. The model appears great but produces a wrong interpretation of results.

```
Call:
lm(formula = Anxiety ~ StressSurvey + Time, data = observDF)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.3904 -0.9896  0.3288  0.6240  2.2912 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)    0.5888     1.0339   0.569   0.5795    
StressSurvey   1.4269     0.1722   8.287 2.62e-06 ***
```

```
Time          -2.7799     1.1111  -2.502   0.0278 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.334 on 12 degrees of freedom
Multiple R-squared:  0.935, Adjusted R-squared:  0.9242
F-statistic: 86.32 on 2 and 12 DF,  p-value: 7.538e-08
```

```
# A tibble: 3 x 5
  term         estimate std.error statistic    p.value
  <chr>           <dbl>     <dbl>     <dbl>      <dbl>
1 (Intercept)     0.589     1.03      0.569 0.580
2 StressSurvey    1.43      0.172     8.29  0.00000262
3 Time           -2.78      1.11     -2.50  0.0278
```

```
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic      p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>        <dbl> <dbl>  <dbl> <dbl> <dbl>
1     0.935         0.924  1.33      86.3 0.0000000754     2  -23.9  55.9  58.7
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
True relationship: Anxiety = Stress + 0.1 × Time

Estimated relationship: Anxiety = 0.5888 + 1.4269 × StressSurvey + -2.7799 × Time

R-squared: 0.935
```

Question 7: Run a multiple regression of Anxiety on both Stress and Time. What would the estimated coefficients be? How would they compare to the true relationship?

When the right predictors are used, the regression estimates a coefficient of 1.0 for Stress and 0.1 for Time, with an R squared of 1.00. That perfectly recovers the underlying data-generating process. The model behaves exactly as expected when the appropriate variables are included.

```
Call:
lm(formula = Anxiety ~ Stress + Time, data = observDF)
```

```
Residuals:
       Min         1Q     Median         3Q        Max 
-1.439e-15 -5.452e-16  4.607e-16  5.639e-16  7.534e-16 

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)    
(Intercept) 9.173e-16  5.931e-16 1.547e+00    0.148    
Stress      1.000e+00  6.700e-17 1.493e+16   <2e-16 ***
Time        1.000e-01  4.718e-16 2.119e+14   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.027e-16 on 12 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1 
F-statistic: 2.549e+32 on 2 and 12 DF,  p-value: < 2.2e-16


# A tibble: 3 x 5
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept) 9.17e-16  5.93e-16   1.55e 0 1.48e-  1
2 Stress      1   e+ 0  6.70e-17   1.49e16 5.51e-189
3 Time        1.00e- 1  4.72e-16   2.12e14 8.20e-167


# A tibble: 1 x 12
  r.squared adj.r.squared    sigma statistic   p.value    df logLik   AIC   BIC
      <dbl>         <dbl>    <dbl>     <dbl>     <dbl> <dbl>  <dbl> <dbl> <dbl>
1         1             1 8.03e-16   2.55e32 1.70e-190     2   502. -996. -
993.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>


True relationship: Anxiety = Stress + 0.1 × Time


Estimated relationship: Anxiety = 0 + 1 × Stress + 0.1 × Time


R-squared: 1
```

Question 8: Compare the R-squared values and coefficient interpretations between the two multiple regression models. Do both models show statistical significance in all of their coefficients? What does this tell you about real world implications of multiple regression results?

Both are statistically significant coefficients with a very high R squared, but they tell different stories. The model employing StressSurvey and Time misrepresents the role of Time, whereas the model that uses Stress and Time identifies the correct effects. This is evidence that statistically significant coefficients and strongly fitting models may not imply correct inference. Selection of variables and their theoretical justification play an important role.

```
# A tibble: 6 x 6
  Model                          term      estimate std.error statistic   p.value
  <chr>                          <chr>        <dbl>     <dbl>     <dbl>     <dbl>
1 Model 3 (StressSurvey + Time) (Interc~  5.89e- 1  1.03e+ 0  5.69e- 1 5.80e-  1
2 Model 3 (StressSurvey + Time) StressS~  1.43e+ 0  1.72e- 1  8.29e+ 0 2.62e-  6
3 Model 3 (StressSurvey + Time) Time     -2.78e+ 0  1.11e+ 0 -2.50e+ 0 2.78e-  2
4 Model 4 (Stress + Time)       (Interc~  9.17e-16  5.93e-16  1.55e+ 0 1.48e-  1
5 Model 4 (Stress + Time)       Stress    1   e+ 0  6.70e-17  1.49e+16 5.51e-
189
6 Model 4 (Stress + Time)       Time      1.00e- 1  4.72e-16  2.12e+14 8.20e-
167


=== R-squared Comparison ===

# A tibble: 2 x 3
  Model                          r.squared adj.r.squared
  <chr>                              <dbl>         <dbl>
1 Model 3 (StressSurvey + Time)      0.935         0.924
2 Model 4 (Stress + Time)            1             1


=== Model Comparison Summary ===

Model 3 (StressSurvey + Time):

  R-squared: 0.935

  Intercept: 0.5888
  StressSurvey coefficient: 1.4269

  Time coefficient: -2.7799


Model 4 (Stress + Time):

  R-squared: 1

  Intercept: 0

  Stress coefficient: 1  (true value = 1.0)

  Time coefficient: 0.1  (true value = 0.1)
```

Question 9: Reflect on Real-World Implications: For each of the two multiple regression models, assume their respective outputs/conclusions were published in academic journals and then subsequently picked up by the popular press. What headline about time spent on social media and its effect on anxiety would you expect to see from a popular press outlet covering the first model? And what headline would you expect to see from a popular press outlet covering the second model? Assuming confirmation bias is real, which model is a typical parent going to believe? Which model will Facebook, Instagram, and TikTok prefer?

- “Study Suggests More Time on Social Media Lowers Anxiety”
- “Increased Social Media Use Linked to Higher Anxiety”

With confirmation bias, parents are more likely to believe the second headline since it confirms pre-existing fears. Social media companies would favor the first headline because it frames usage as something positive. Both of these interpretations arise from the same regression output and demonstrate how easily statistical results can be misrepresented.

Question 10: Avoiding Misleading Statistical Significance: Reflect on this tip to avoid being misled by statistically significant results: splitting the sample into meaningful subsets (“statistical regimes”), and using graphical diagnostics for linearity rather than blind reliance on “canned” regressions: Apply this approach to multiple regression of Anxiety on both StressSurvey and Time by analyzing a smartly chosen subset of the data. What specific subset did you choose and why? Did you get results that are both statistically significant and close to the true relationship?

I restricted the sample to observations with StressSurvey   6 and reestimated the model. In this subset, the coefficient on Time was closer to its true value (+ 0.1) and no longer reversed sign. The data also appeared more linear. This exercise shows that partitioning the data into meaningful regimes and examining diagnostic plots can prevent erroneous conclusions that arise from blind trust in a single regression output.