# 22401 HW4

Bin Yu

Feb 14, 2025

## Question 1

### (a)

```
. regress ozone rad temp wind
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 11.3350041 | 3 | 3.77833471 |
| Residual | 5.33014388 | 26 | .205005534 |
| Total | 16.665148 | 29 | .574660276 |

| | | |
|---|---|---|
| Number of obs | = | 30 |
| F(3, 26) | = | 18.43 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.6802 |
| Adj R-squared | = | 0.6433 |
| Root MSE | = | .45278 |

| ozone | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| rad | .0032856 | .001004 | 3.27 | 0.003 | .0012218 | .0053493 |
| temp | .0473549 | .0121477 | 3.90 | 0.001 | .0223849 | .0723249 |
| wind | -.0696736 | .0274474 | -2.54 | 0.017 | -.1260926 | -.0132546 |
| _cons | -.343442 | 1.0589 | -0.32 | 0.748 | -2.520041 | 1.833157 |

Figure 1: Regression Results

**Interpretation**:

- **Coefficient and significance:**
  - `rad`: The coefficient is positive and statistically significant ($p < 0.01$). This suggests that as solar radiation (`rad`) increases, `ozone` levels tend to rise. Specifically, every 1-unit increase in `rad` is associated with an increase of approximately 0.0033 in `ozone`, holding the other variables constant.
  - `temp`: The coefficient is also positive and significant ($p < 0.01$), indicating that higher temperatures (`temp`) are linked to higher `ozone` levels. A 1-degree increase in temperature corresponds to an increase of about 0.0474 in `ozone`, holding the other variables constant.
  - `wind`: The coefficient is negative and statistically significant ($p < 0.05$), suggesting that stronger wind (`wind`) is associated with lower `ozone`. A 1-unit increase in wind speed reduces `ozone` by roughly 0.07, holding the other variables constant.
- **Model fit:** The $R^2$ is about 0.68, meaning the model explains approximately 68% of the variation in `ozone`.

Using the following code to compute the residuals and predicted values.

```
regress ozone rad temp wind

predict yhat, xb
predict rawres, resid

twoway scatter rawres ozone,                    ///
    title("Raw Residuals vs. Actual Ozone")     ///
    xtitle("Actual Ozone")                      ///
    ytitle("Raw Residuals")
```

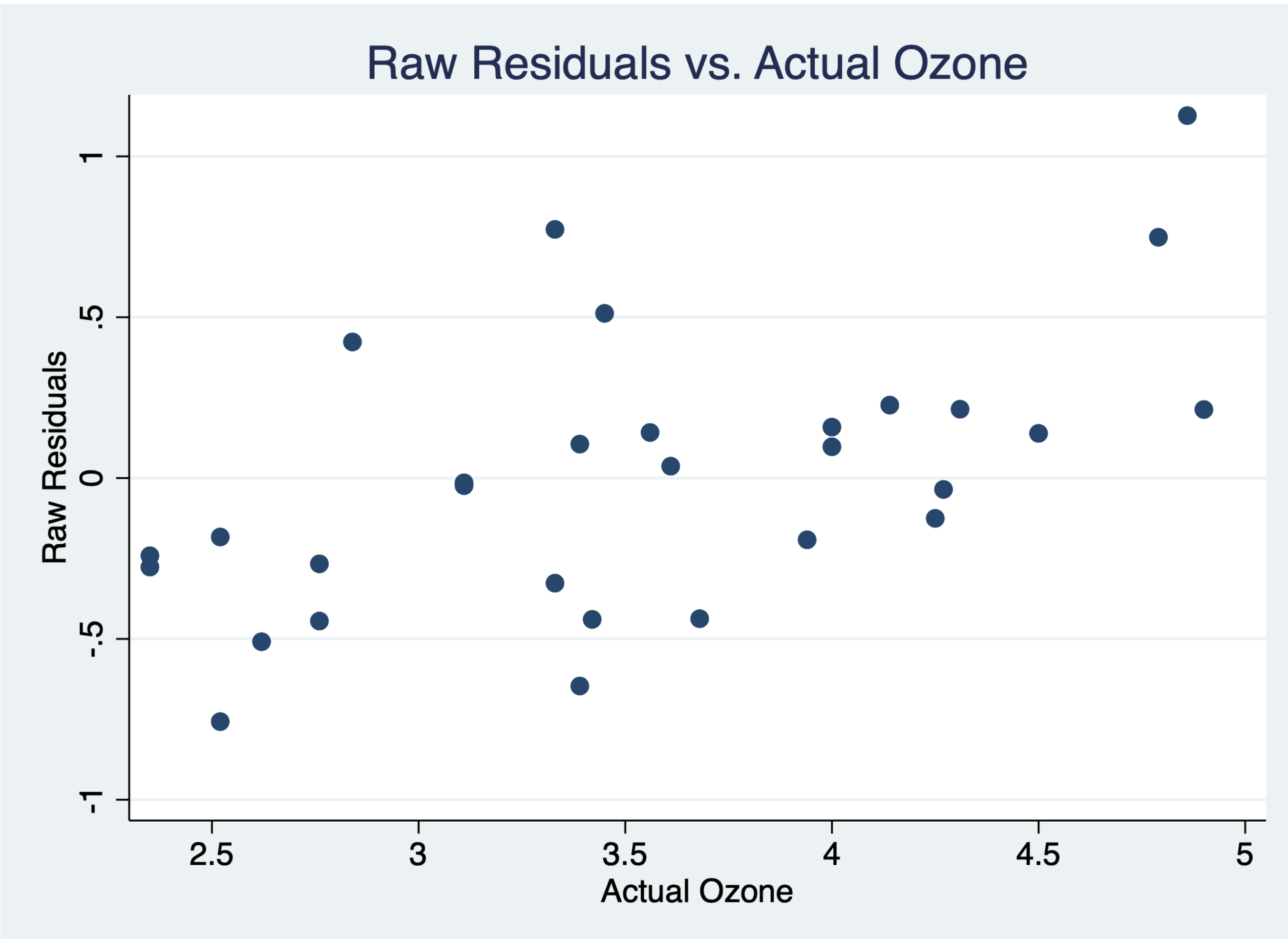



Figure 2: Raw Residuals vs. Actual Ozone

**Interpretation**

- **Distribution Shift (Negative on the Left, Positive on the Right):** Observing the plot, many residuals are negative for lower ozone values (left side) and positive for higher ozone values (right side). This could indicate a potential nonlinearity in the relationship, where the model might be under-predicting at higher ozone levels and over-predicting at lower ozone levels. Although not extremely severe, it is worth further investigation (e.g., by checking polynomial terms or transformations).
- **Overall Randomness and Magnitude:** Despite the left/right tendency, within each region the residuals appear relatively scattered at random. The points in the scatter plot do not follow a clear pattern or trend

with respect to the actual ozone values. This suggests that, broadly, the linear model form is appropriate and there is no obvious sign of systematic curvature or major violation of the linearity assumption.

- **Constant Variance (Homoscedasticity):** There is no obvious "fanning out" or narrowing that would strongly suggest heteroskedasticity. The spread of residuals from the horizontal axis is roughly consistent across the range of ozone.

## (b)

**Stata Code**

```
predict rstd, rstandard
twoway scatter rstd yhat,                          ///
    mlabel(id)                                      ///
    yscale(range(-3 4))                             ///
    title("Standardized Residuals vs. Fitted Values")  ///
    xtitle("Fitted Ozone (yhat)")                   ///
    ytitle("Standardized Residuals")                ///
    yline(-2 2, lstyle(dash))                       ///
    legend(off)
```

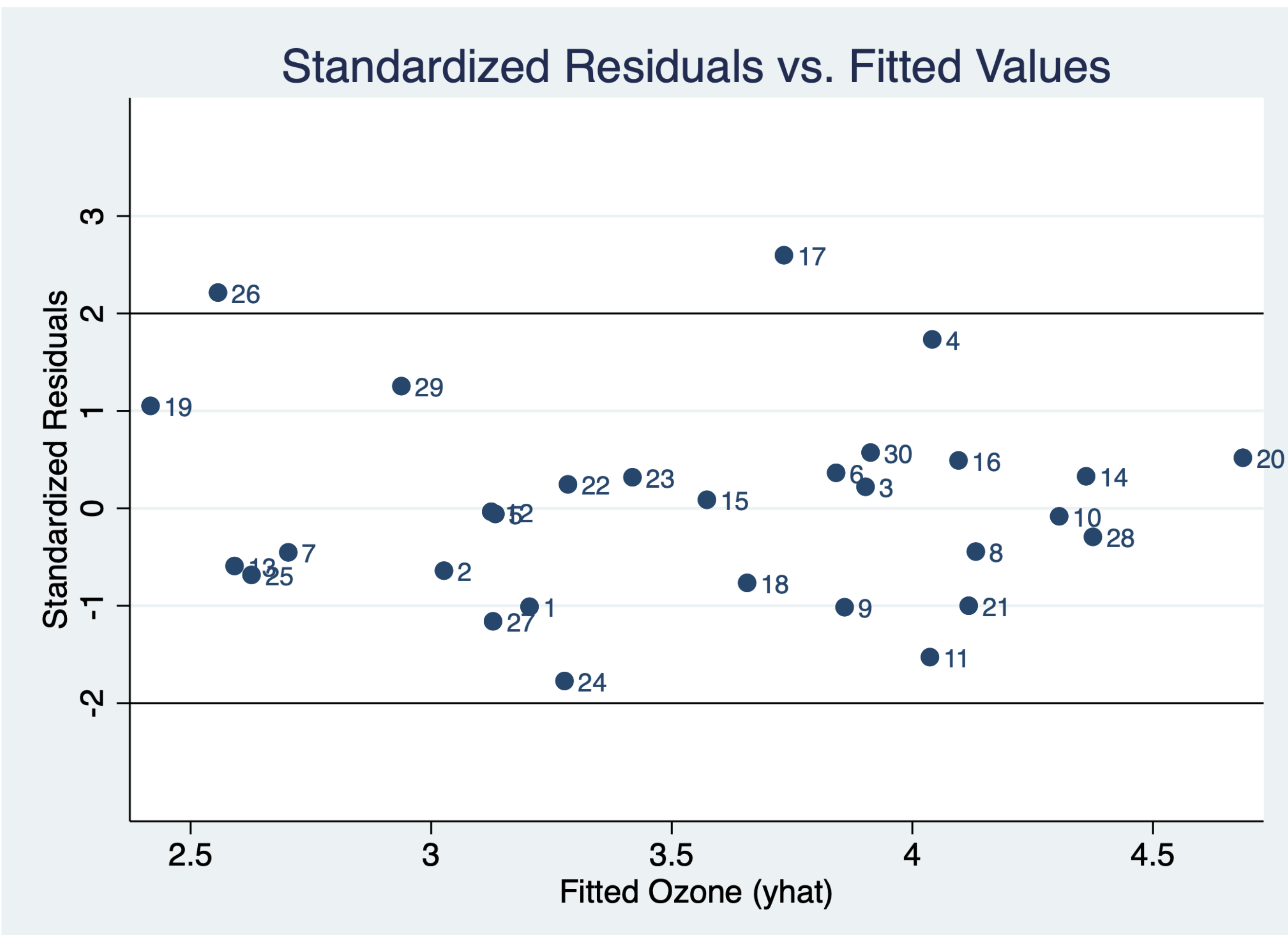



Figure 3: Standardized Residuals vs. Fitted Values

Figure 5 shows the standardized residuals versus fitted values, providing a check for homoscedasticity and extreme points. The observations with id 17 and 26 have absolute standardized residuals greater than 2, indicating potential outliers.

## (c)

### Stata Code

```
list id ozone rad temp wind rstd if abs(rstd) > 2
```

```
. list id ozone rad temp wind rstd if abs(rstd) > 2
```

| | id | ozone | rad | temp | wind | rstd |
|---|---|---|---|---|---|---|
| 17. | 17 | 4.86 | 223 | 79 | 5.7 | 2.597539 |
| 26. | 26 | 3.33 | 284 | 72 | 20.7 | 2.214397 |

Figure 4: Standardized residuals with absolute standardized residuals greater than 2

Observations 17 and 26 have standardized residuals ($|rstd| > 2$) and are flagged as potential outliers. Below are summary statistics comparing these outliers to the remaining observations.

```
. sum ozone rad temp wind if inlist(id, 17, 26)
. sum ozone rad temp wind if !inlist(id, 17, 26)

. sum rawres rstd yhat ozone if inlist(id, 17, 26)
. sum rawres rstd yhat ozone if !inlist(id, 17, 26)
```

```
.
. sum ozone rad temp wind if inlist(id, 17, 26)

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       ozone |          2       4.095    1.081874       3.33       4.86
         rad |          2       253.5    43.13351        223        284
        temp |          2        75.5    4.949747         72         79
        wind |          2        13.2     10.6066        5.7       20.7

.
. sum ozone rad temp wind if !inlist(id, 17, 26)

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       ozone |         28    3.495357    .7409578       2.35        4.9
         rad |         28    202.1786     94.5935          7        323
        temp |         28       80.75    8.280387         64         94
        wind |         28    8.346429    2.983595        2.3       15.5

.
. sum rawres rstd yhat ozone if inlist(id, 17, 26)

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      rawres |          2    .9499483    .2501968   .7730325   1.126864
        rstd |          2    2.405968    .2709223   2.214397   2.597539
        yhat |          2    3.145052    .8316767   2.556967   3.733136
       ozone |          2       4.095    1.081874       3.33       4.86

. sum rawres rstd yhat ozone if !inlist(id, 17, 26)

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      rawres |         28   -.0678535    .3513902  -.7570625   .7487786
        rstd |         28   -.1550411     .827566  -1.772312   1.733872
        yhat |         28    3.563211    .6181489   2.416788   4.686921
       ozone |         28    3.495357    .7409578       2.35        4.9
```

Figure 5: Comparisons between outliers and others

**Findings**

- **Higher Ozone and Radiation on Average:** The two outliers have an average ozone of 4.095 (vs. 3.495 among the others) and a higher average solar radiation (253.5 vs. 202.18). This suggests they lie in conditions more conducive to elevated ozone levels.
- **Lower Temperature but Higher Wind:** Outliers' mean temperature (75.5) is lower than that of the other group (80.75), their wind speed is substantially higher (13.2 vs. 8.35). This combination might deviate from the typical pattern captured by the model.
- **Model Under-Prediction:** The outliers show a mean `rawres` of about +0.95, whereas the non-outliers average around −0.07. Likewise, the `rstd` for these outliers is around +2.40, confirming that the model underestimates ozone for these two points, their fitted values (`yhat` $\approx$ 3.15) are notably below the actual

ozone (4.095), the model may require additional factors or a different functional form to accurately capture the high-wind, moderate-temperature regime in which these observations occur.

Overall, observations 17 and 26 differ from the rest of the sample by combining relatively high ozone levels with higher wind and lower temperature. The linear model under-predicts their ozone, reflected by large positive (and standardized) residuals. Additional predictors or a nuanced understanding of meteorological conditions might be needed to account for these outliers.

## (d)

**Stata Code**

```
qnorm rstd
summ rstd
```

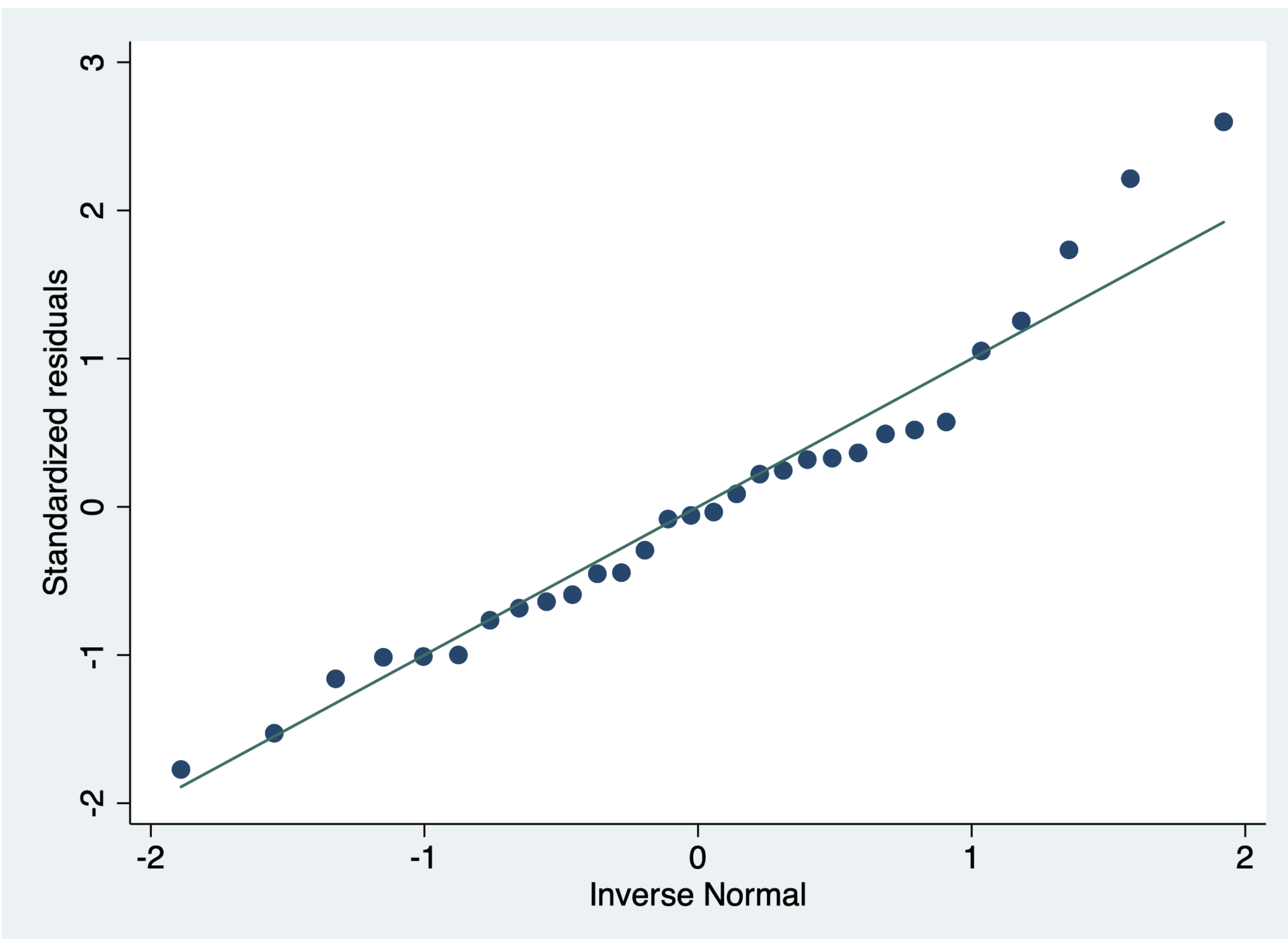



Figure 6: Q-Q Plot of Standardized Residuals

```
. summ rstd
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| rstd | 30 | .0156928 | 1.030698 | -1.772312 | 2.597539 |

Figure 7: Mean and Variance of Standardized Residuals

**Interpretation of Q-Q Plot and Summary Statistics**

- **General Conformity to Normality:** The Q-Q plot (Figure 24) shows that most points lie close to the straight line, indicating that the standardized residuals largely follow a normal distribution.
- **Mean and Standard Deviation:** The mean of the standardized residuals is approximately 0.016 and the standard deviation is around 1.03, both of which are close to the theoretical values of 0 and 1, respectively. This supports the normality assumption.
- **High-End Tail:** The highest standardized residual is about 2.60. This suggests that while the upper tail is a bit heavier, it does not overwhelmingly violate the normality assumption.
- **Overall Conclusion:** Given the Q-Q plot alignment and near-ideal summary statistics, there is no strong evidence of non-normal errors. Minor deviations in the upper tail (e.g., a few points above the line) can be monitored, but do not appear to invalidate the assumption of normality for this model's residuals.

## (e)

## Stata Code

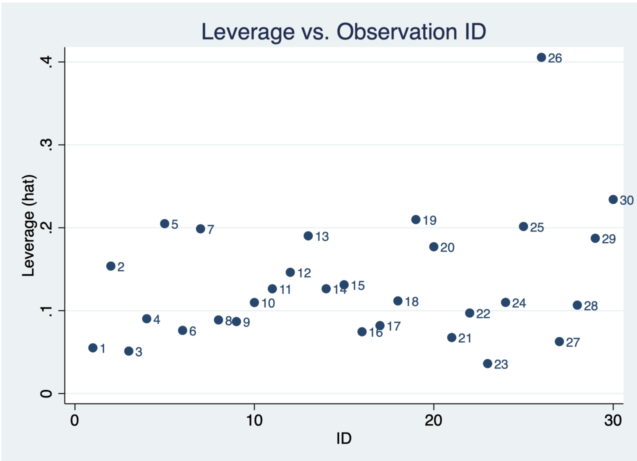
```
predict leverage, hat
predict cookd, cooksd

twoway scatter leverage id, mlabel(id) ///
    title("Leverage vs. Observation ID") ///
    xtitle("ID") ///
    ytitle("Leverage (hat)")

twoway scatter cookd id, mlabel(id) ///
    title("Cook's Distance vs. Observation ID") ///
    xtitle("ID") ///
    ytitle("Cook's Distance")

list id ozone rstd cookd leverage if abs(rstd)>2
```

### Plots and Summary Statistics



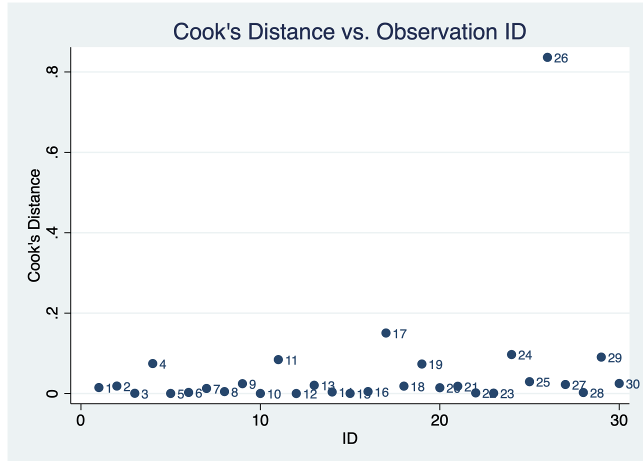




Figure 8: Leverage vs. Observation ID (left) and Cook's Distance vs. Observation ID (right)

```
. sum leverage,detail

                          Leverage
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .0361229       .0361229
 5%     .0511926       .0511926
10%     .0589215       .0551465       Obs                  30
25%     .0819771       .0626964       Sum of Wgt.          30

50%     .1108429                      Mean           .1333333
                        Largest       Std. Dev.      .0750839
75%     .1873134       .2049186
90%     .2073998       .2098809       Variance       .0056376
95%      .234114        .234114       Skewness       1.620502
99%     .4055449       .4055449       Kurtosis       6.709178

.
. predict cookd, cooksd

. sum cookd,detail

                          Cook's D
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .0000533       .0000533
 5%     .0002085       .0002085
10%     .0002535       .0002164       Obs                  30
25%     .0025574       .0002906       Sum of Wgt.          30

50%     .0164936                      Mean           .0549726
                        Largest       Std. Dev.      .1522104
75%     .0294941       .0906614
90%     .0938307       .0969999       Variance        .023168
95%     .1506271       .1506271       Skewness        4.73156
99%     .8363168       .8363168       Kurtosis       24.78304

. list id ozone rstd cookd leverage if abs(rstd)>2
```

| | id | ozone | rstd | cookd | leverage |
|---|---|---|---|---|---|
| 17. | 17 | 4.86 | 2.597539 | .1506271 | .0819771 |
| 26. | 26 | 3.33 | 2.214397 | .8363168 | .4055449 |

Figure 9: Summary Statistics

**Leverage (hat) values:**

- *Mean Leverage*: 0.1333, with a standard deviation of 0.0751.
- *Max Leverage*: Observation 26 exhibits the highest leverage ($\approx$ 0.4055), considerably above the rest of the

sample.

- Large leverage values suggest that the corresponding observation has predictor values substantially different from the majority of data and can exert a strong pull on the fitted regression line.

**Cook's Distance:**

- *Mean Cook's D*: 0.055, with a standard deviation of 0.1522.
- *Max Cook's D*: Observation 26 stands out again with a Cook's distance around 0.8363, which is far higher than the next largest value.
- #26 may have considerable influence on the regression results.

**Specific Observations (list if $|rstd| > 2$):**

- `id = 17`: Standardized residual = 2.60, Cook's D = 0.1506, leverge = 0.0820
- `id = 26`: Standardized residual = 2.21, Cook's D = 0.8363, leverage = 0.4055

**Interpretation**

- **Observation 26:** Has both high leverage and high Cook's distance, indicating that it is very influential. Removing or altering this data point could substantially change the fitted model.
- **Observation 17:** Displays a large standardized residual but less extreme Cook's distance and leverage. This suggests it is an outlier in terms of vertical distance from the regression line, though not as powerful an influencer of overall model fit.

# Question 2

## (a)

**Model Specification:** We regress `Sales` on three predictors:

$$\text{Sales} = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Income} + \beta_3 \cdot \text{Price} + \varepsilon.$$

```
. regress Sales Age Income Price
```

| Source | SS | df | MS | | | |
|---|---|---|---|---|---|---|
| Model | 15594.4257 | 3 | 5198.1419 | Number of obs | = | 51 |
| Residual | 35831.0197 | 47 | 762.362122 | F(3, 47) | = | 6.82 |
| | | | | Prob > F | = | 0.0007 |
| | | | | R-squared | = | 0.3032 |
| | | | | Adj R-squared | = | 0.2588 |
| Total | 51425.4454 | 50 | 1028.50891 | Root MSE | = | 27.611 |

| Sales | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| Age | 4.155908 | 2.198699 | 1.89 | 0.065 | -.2673039 | 8.579119 |
| Income | .019281 | .0068833 | 2.80 | 0.007 | .0054337 | .0331284 |
| Price | -3.399234 | .9891719 | -3.44 | 0.001 | -5.389191 | -1.409277 |
| _cons | 64.24826 | 61.93301 | 1.04 | 0.305 | -60.34488 | 188.8414 |

Figure 10: Summary Statistics

**Results:**

- **Overall fit:** The $R^2$ is about 0.30, indicating that around 30% of the variation in `Sales` is explained by `Age`, `Income`, and `Price`. The F-test is significant ($p = 0.0007$), suggesting the model as a whole has predictive power.
- **Coefficients:**
  - `Age` has a positive coefficient ($\hat{\beta}_1 \approx 4.16$) but is only marginally significant ($p = 0.065$).
  - `Income` is positively associated with `Sales` ($p = 0.007$).
  - `Price` has a negative coefficient ($p = 0.001$).

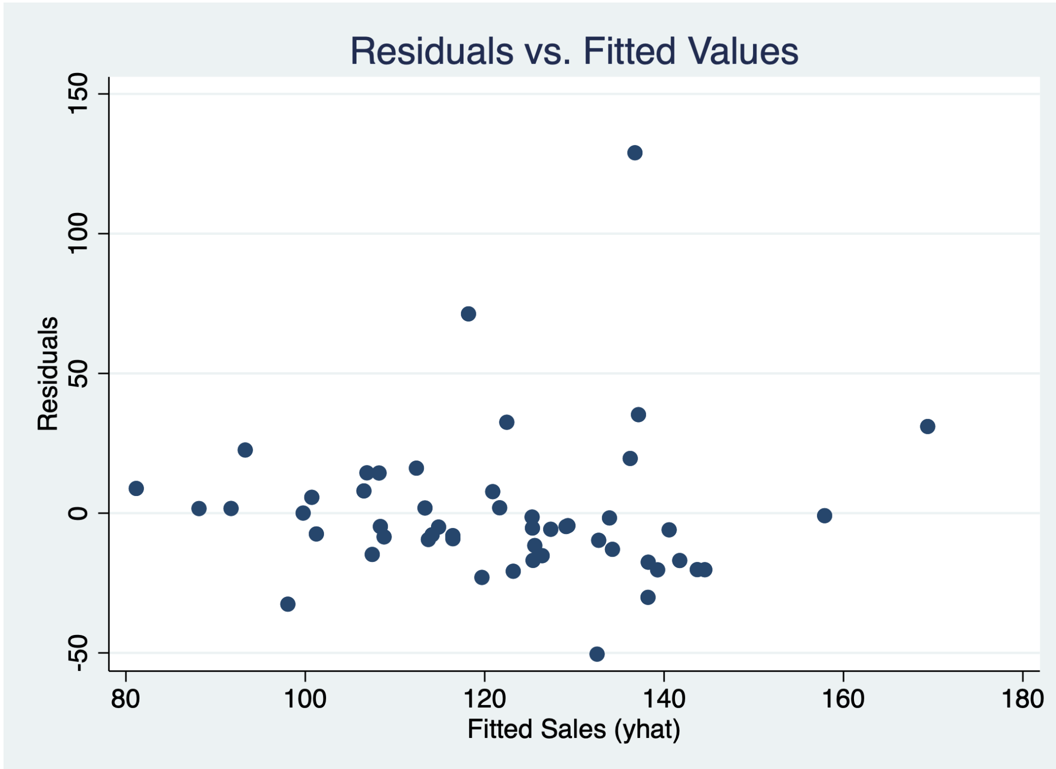



Figure 11: Residuals vs. Fitted Values

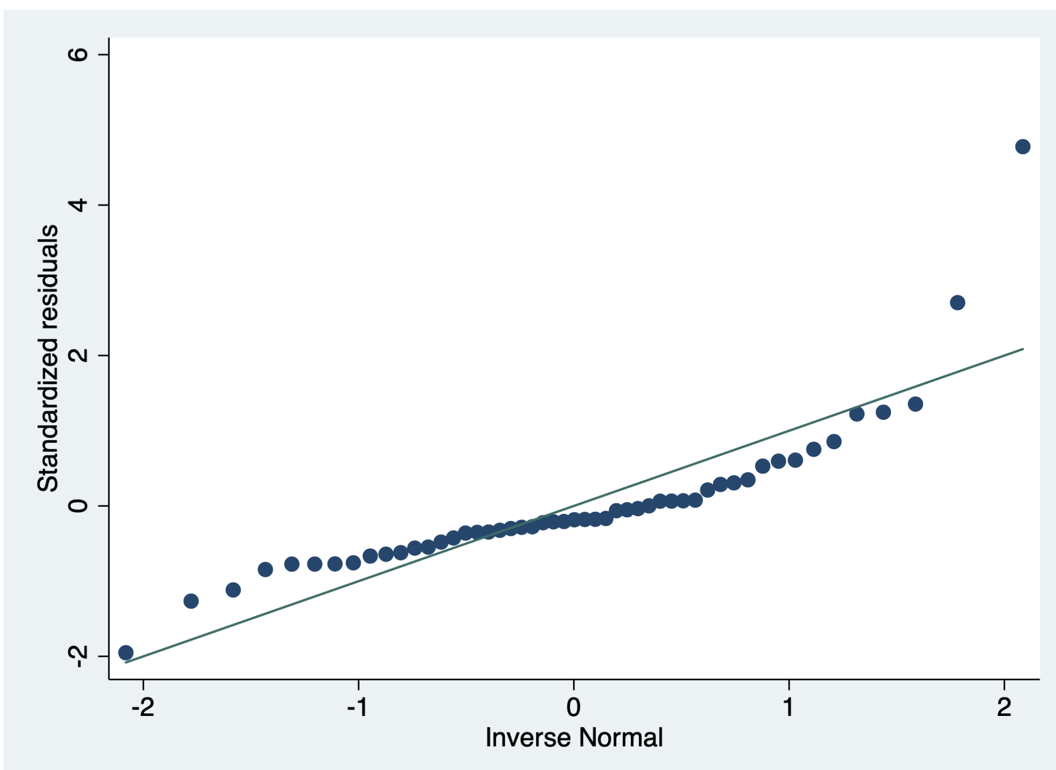



Figure 12: Q-Q Plot of Standardized Residuals

```
. summm rstd
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| rstd | 51 | .00255 | 1.006584 | -1.951511 | 4.776447 |

Figure 13: Summary of Standardized Residuals

**Residual Analysis:**

- **Residuals vs. Fitted Values (Figure 11):**
  - The scatter of residuals around the zero line does not exhibit a clear "U-shape" or other strong pattern, suggesting no major violation of linearity.
  - While most points cluster within $\pm 50$, there are one or two observations exceeding these bounds (e.g., a point above $+100$), which may indicate unusual consumption levels relative to what the model predicts.
  - The variability of residuals appears roughly consistent across the range of fitted values, suggesting no severe heteroskedasticity.
- **Normality Check:**
  - *Q-Q Plot (Figure 24):* Most standardized residuals lie close to the 45-degree reference line, indicating that the distribution of errors is reasonably normal for the bulk of observations. However, a small cluster of points deviate near the upper tail (above 2.5), with one point extending beyond 4.7. This outlier suggests the possibility of a right-skewed tail or a single data point that does not fit well under the current model.
  - *Summary Statistics:* The mean standardized residual (0.0026) is nearly zero, and the standard deviation (1.0066) is close to 1, which is consistent with normally distributed errors, so the assumption of normal residuals is largely met.

## (b)

To diagnose whether specific ranges of each predictor are associated with unusually large residuals, we plotted the standardized residuals (y-axis) against each predictor (x-axis): `Age`, `Income`, and `Price`.

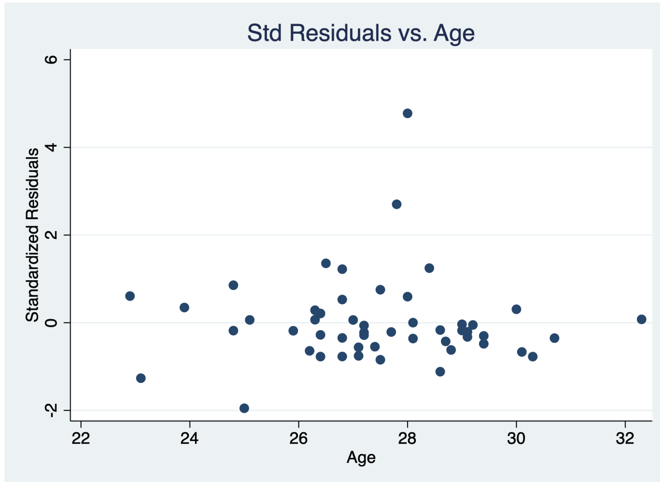



Figure 14: Std Residuals vs. Age

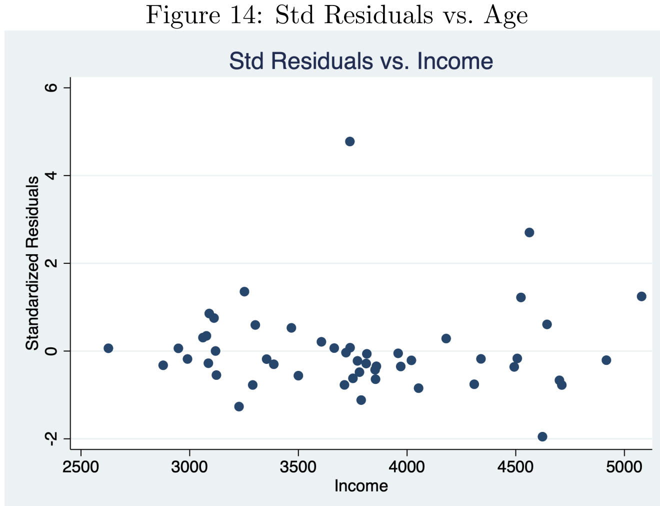



Figure 15: Std Residuals vs. Income


Figure 16: Std Residuals vs. Price

**Interpretation**

- **Std Residuals vs. Age (Figure 14):**
  - Residuals appear scattered randomly across the range of `Age` values (roughly 22 to 32).
  - One or two observations show much larger positive residuals (exceeding +4) but do not strictly occur at the extremes of `Age`.
  - Overall, there is no obvious trend such as increasing or decreasing residual magnitude with `Age`.
- **Std Residuals vs. Income (Figure 15):**
  - The majority of points lie between $-2$ and $+2$, with one apparent outlier near `Income` $\approx 4000$ displaying a standardized residual above 4.
  - At the low and high ends of `Income` (below 3000 or above 4500), the residuals remain within a moderate range. However, as the value of income goes up, the extreme residual seems occur more, suggesting some potential pattern of misfit at extreme values of `Income`.
- **Std Residuals vs. Price (Figure 16):**
  - Residuals cluster randomly near 0 for most observations, though a couple of points exceed $\pm 3$.
  - There's no clear evidence that extreme `Price` values (e.g., above 40 or below 30) systematically produce large residuals.
  - A single observation near `Price` $\approx 34$ stands out with a standardized residual around $+5$, which could be a potential outlier.

Overall, while the majority of observations remain within $|rstd| < 2$, one observation with standardized residual that exceed $+4$ do not consistently occur at the far ends of `Age`, `Income`, or `Price`. It occur around the median of `Age`, `Income`, or `Price`. This suggests that *some* outliers may reflect other factors not captured by these predictors.

## (c)

**Identification of Outliers Using Standardized Residuals.**
We flagged any observation with |`rstd`| $> 2$ as a potential outlier. The table below shows two cases:

```
. list state age income price sales rstd if abs(rstd) > 2

     +-------------------------------------------------+
     | state   age   income   price   sales       rstd |
     |-------------------------------------------------|
 29. |   NV    27.8    4563      44   189.5   2.702517   |
 30. |   NH      28    3737    34.1   265.7   4.776447   |
     +-------------------------------------------------+
```

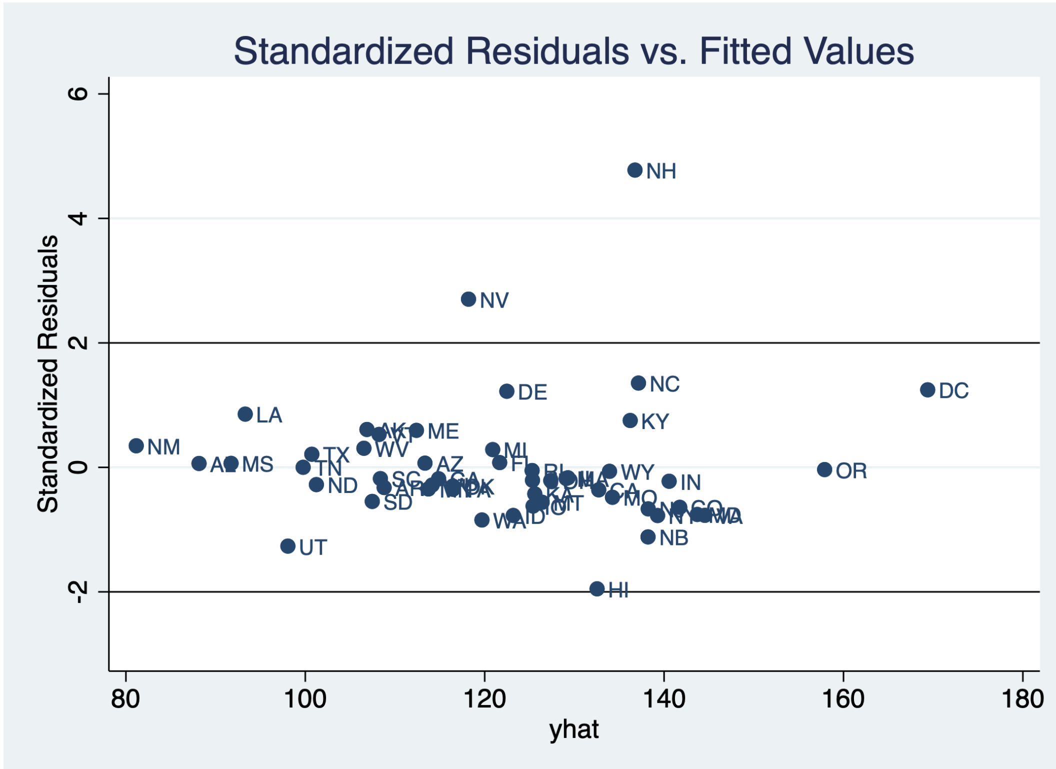



Figure 17: Standaralized Residual vs. fitted values

- **Nevada (NV)** has a standardized residual of 2.70 and predicted sales about 189.5. It is moderately beyond the $\pm 2$ threshold.
- **New Hampshire (NH)** has a standardized residual near 4.78, suggesting a substantial gap between its actual and predicted sales.

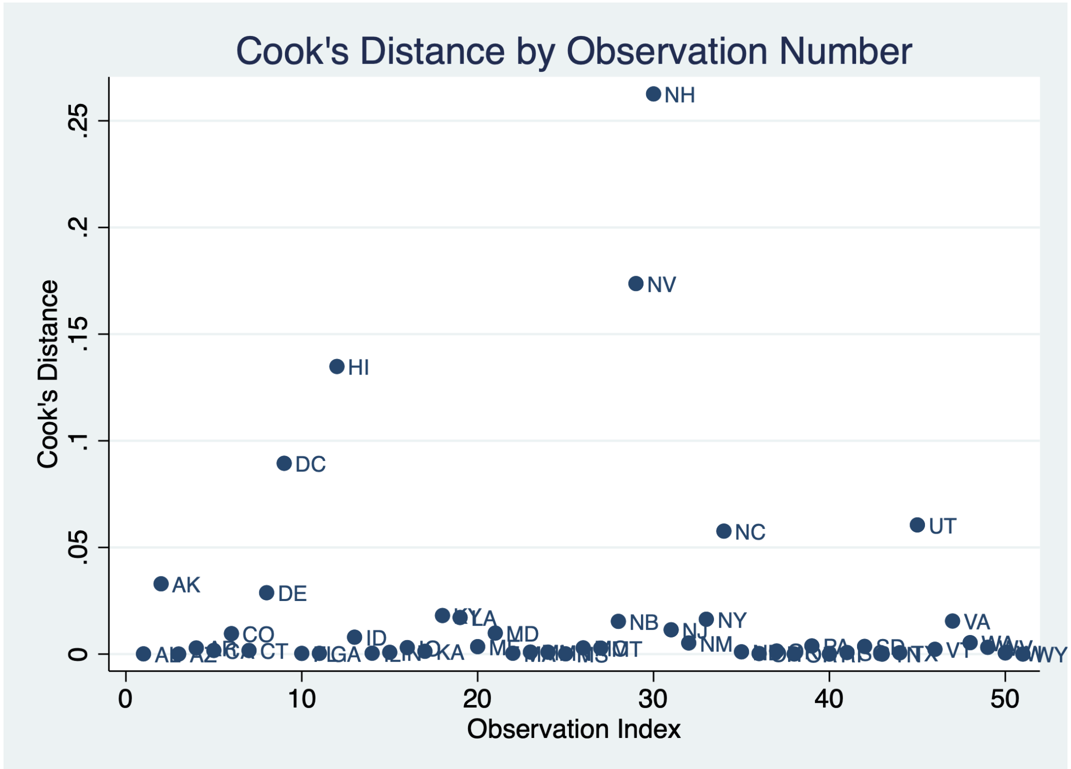



Figure 18: Cook's Distance by Observation Number

**Cook's Distance:**

- The Cook distance quantifies how much removing a single observation would change the overall regression estimates (coefficients).
- Observations of `NH` and `NV` show comparatively high Cook's D, indicating that dropping either one could meaningfully alter the slope estimates.

**Comparisons to the Rest of the Dataset.**

**Summary of Key Predictors and Response**
The table provides descriptive statistics for `Sales`, `Age`, `Income`, and `Price`, grouped by (a) all states except NV and NH (b) NV only, and (c) NH only:

```
. sum sales age income price if state!="NV" & state!="NH"

  Variable | Obs   Mean     Std. Dev.   Min     Max
  ---------+-----------------------------------------
     sales | 49  117.21       22.87    65.50  200.40
       age | 49   27.45        1.91    22.90   32.30
    income | 49 3747.94      595.69  2626.00 5079.00
     price | 49   38.03        4.09    29.00   45.50

. sum sales age income price if state=="NV"
```

```
Variable | Obs   Mean     Std. Dev.   Min    Max
---------+---------------------------------------
   sales |  1  189.50          .    189.5  189.50
     age |  1   27.80          .    27.80   27.80
  income |  1  4563.00         .  4563.00 4563.00
   price |  1   44.00          .    44.00   44.00

. sum sales age income price if state=="NH"

Variable | Obs   Mean     Std. Dev.   Min    Max
---------+---------------------------------------
   sales |  1  265.70          .   265.70  265.70
     age |  1   28.00          .    28.00   28.00
  income |  1  3737.00         .  3737.00 3737.00
   price |  1   34.10          .    34.10   34.10
```

**Observations:**

- *Nevada (NV)* has relatively high `Sales` (189.5) compared to the non-outlier mean (117.2) and a higher-than-average `Income` (4563) to the non-outlier mean(3747.94). Other features of NV are similar to non-outliers.
- *New Hampshire (NH)* has Sales = 265.7, far exceeding the mean of non-outliers (117.2). Other features of NH are similar to non-outliers.

**Fitted Values and Residuals**

After regenerating the fitted values (`yhat`), ordinary residuals (`rawresid`), and standardized residuals (`i_stresid`, `e_stresid`), we compare non-outliers vs. NV vs. NH:

```
. sum sales yhat rawresid i_stresid e_stresid if state!="NV" & state!="NH"

Variable     | Obs   Mean      Std. Dev.    Min        Max
-------------+----------------------------------------------
 sales       | 49  117.2122    22.87057    65.5       200.4
 yhat        | 49  121.299     17.88211    81.15414   169.3894
 rawresid    | 49  -4.08676    16.64644   -50.4303    35.25572
 i_stresid   | 49  -0.14998    0.63641    -1.95151    1.35471
 e_stresid   | 49  -0.14994    0.64029    -2.01396    1.36718

. sum sales yhat rawresid i_stresid e_stresid if state=="NV"

Variable     | Obs   Mean      Std. Dev.    Min        Max
-------------+----------------------------------------------
 sales       |  1  189.5           .        189.5      189.5
 yhat        |  1  118.1956        .        118.1956   118.1956
 rawresid    |  1   71.30443       .         71.30443   71.30443
 i_stresid   |  1    2.702517      .          2.702517   2.702517
 e_stresid   |  1    2.909188      .          2.909188   2.909188

. sum sales yhat rawresid i_stresid e_stresid if state=="NH"

Variable     | Obs   Mean      Std. Dev.    Min        Max
-------------+----------------------------------------------
 sales       |  1  265.7           .        265.7      265.7
 yhat        |  1  136.753         .        136.753    136.753
```

```
rawresid    |  1   128.947          .       128.947    128.947
i_stresid   |  1     4.776447       .         4.776447   4.776447
e_stresid   |  1     6.587276       .         6.587276   6.587276
```

**Interpretation of Residuals**

- **Non-Outliers**:
  - The mean raw residual is around $-4.09$, with a standard deviation of about 16.65.
  - Standardized residuals remain within about $(-2.01, +1.37)$, indicating that no severe outliers lie in this main group.
- **Nevada (NV)**:
  - The model under-predicts `Sales` by roughly 71.3 units, with $i\_stresid \approx 2.70$ and $e\_stresid \approx 2.91$—both beyond the usual $|2|$ cutoff.
  - This confirms NV is an outlier in terms of *vertical distance*.
- **New Hampshire (NH)**:
  - The gap between actual and predicted `Sales` is even larger (+128.95), leading to $i\_stresid \approx 4.78$ and $e\_stresid \approx 6.59$.
  - These values far exceed typical thresholds for standardized residuals, making NH a clear outlier.

**Commentary**

- **Magnitude of Outliers:** Nevada is beyond the common $|2|$ boundary, and New Hampshire is far beyond, indicating it contributes unusual variability that might heavily influence regression coefficients or predictions.
- **Possible Explanation:**
  - NV and NH's other demographic factors could drive elevated cigarette consumption not captured by `Age`, `Income`, and `Price` alone.
- **Implications:**
  - Therefore, since NH and NV's sales cannot be explained by extreme `Age`, `Income`, and `Price`, including additional predictors may reduce the impact of these points.
  - Alternatively, if NV and NH represent genuine but exceptional cases, one could model them separately (e.g., using dummy variables) to avoid distorting the primary relationships in the rest of the data.

# Question 3

## (a)

**Model Specification:** We fit a linear model using the natural logarithms of brain weight (ln(brainwt)) and body weight (ln(bodywt)), plus an indicator for being a primate (`primate`):

$$\ln(\text{brainwt}) \;=\; \beta_0 \;+\; \beta_1 \ln(\text{bodywt}) \;+\; \beta_2\, \text{primate} \;+\; \varepsilon.$$

Using the following code in Stata to fit the log–log model that includes `primate` as a binary indicator and perform the F-test:

```
 gen byte primate = 0

. replace primate = 1 if inlist(name, "Gorilla", "Human", "Chimpanzee", "Rhesus monkey", "Potar monkey")

regress logbrainwt logbodywt primate
. test primate
```

```
. regress logbrainwt logbodywt primate

      Source |       SS           df       MS      Number of obs   =        28
-------------+----------------------------------   F(2, 25)        =     29.21
       Model |  108.851548         2   54.425774   Prob > F        =    0.0000
    Residual |  46.5754685        25  1.86301874   R-squared       =    0.7003
-------------+----------------------------------   Adj R-squared   =    0.6764
       Total |  155.427017        27  5.75655617   Root MSE        =    1.3649

------------------------------------------------------------------------------
  logbrainwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   logbodywt |   .5019648   .0696973     7.20   0.000     .3584205    .6455091
     primate |   1.874158   .6738213     2.78   0.010     .4863966    3.261919
       _cons |   2.197712   .3899392     5.64   0.000     1.394617    3.000807
------------------------------------------------------------------------------
```

Figure 19: Regression Results

**F-test for the Primate Indicator**

```
 ( 1)  primate = 0

       F(  1,    25) =    7.74
            Prob > F =    0.0101
```

This shows that $\beta_{\text{primate}} \neq 0$ is significant at the 1% level ($p = 0.0101$), thus rejecting the null hypothesis that the primate indicator has no effect, which means that primate is a significant predictor.

**Comparing Models: With vs. Without Primate.** We fit two regressions: one omitting `primate`, and one including it.

```
. regress logbrainwt logbodywt

      Source |       SS           df       MS      Number of obs   =        28
-------------+----------------------------------   F(1, 26)        =     40.26
       Model |  94.4390272         1  94.4390272   Prob > F        =    0.0000
    Residual |  60.9879893        26   2.3456919   R-squared       =    0.6076
-------------+----------------------------------   Adj R-squared   =    0.5925
       Total |  155.427017        27  5.75655617   Root MSE        =    1.5316

------------------------------------------------------------------------------
  logbrainwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   logbodywt |   .4959947   .0781694     6.35   0.000     .3353152    .6566742
       _cons |   2.554898    .413137     6.18   0.000     1.705683    3.404113
------------------------------------------------------------------------------
```

Figure 20: Regression Results without primate

**Interpretation.**

- **Model Fit.** Including `primate` increases $R^2$ from 0.608 to 0.700, and the adjusted-$R^2$also increases. This indicates that primate status explains additional variability in ln(brainwt) beyond body weight alone.
- **Statistical Significance.** The F-test yields $F = 7.74$ with $p = 0.0101$, confirming that $\beta_{\text{primate}}$ is significantly different from zero at about the 1% level.
- **Primate Effect.** The estimated coefficient of 1.874 on `primate` implies that, holding ln(bodywt) constant, primates have an average brain weight $\exp(1.874) \approx 6.52$ times larger than non-primates.

## (b)

**Model Specification:** use the log–log model by including an interaction term,

$$\ln(\text{brainwt}) \;=\; \beta_0 \;+\; \beta_1 \ln(\text{bodywt}) \;+\; \beta_2\,\text{primate} \;+\; \beta_3\big[\ln(\text{bodywt}) \times \text{primate}\big] \;+\; \varepsilon.$$

Using the following code:

```
. gen interaction = logbodywt * primate
. regress logbrainwt logbodywt primate interaction
```

```
. regress logbrainwt logbodywt primate interaction
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 108.868136 | 3 | 36.2893787 |
| Residual | 46.5588804 | 24 | 1.93995335 |
| Total | 155.427017 | 27 | 5.75655617 |

| | | |
|---|---|---|
| Number of obs | = | 28 |
| F(3, 24) | = | 18.71 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.7004 |
| Adj R-squared | = | 0.6630 |
| Root MSE | = | 1.3928 |

| logbrainwt | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| logbodywt | .5029183 | .0718654 | 7.00 | 0.000 | .3545954 | .6512411 |
| primate | 2.03779 | 1.898455 | 1.07 | 0.294 | -1.880428 | 5.956007 |
| interaction | -.0463171 | .5008855 | -0.09 | 0.927 | -1.080094 | .9874599 |
| _cons | 2.194066 | .3998585 | 5.49 | 0.000 | 1.368798 | 3.019333 |

Figure 21: Regression Results

The `test interaction` command yields:

```
. test interaction

 ( 1)  interaction = 0

       F(  1,    24) =    0.01
            Prob > F =    0.9271
```

indicating that $\beta_3$ is statistically insignificant.

**Interpretation** **No Significant Slope Difference**: The interaction coefficient ($\beta_3 \approx -0.046$, $p = 0.93$) implies that primates do not have a systematically different slope relating ln(bodywt) to ln(brainwt). In other words, the rate at which brain weight changes with body weight on log-scale is the same for primates and non-primates.

**Conclusion:** While the `primate` indicator alone improved the model in Part (a), adding an interaction term does not further clarify the relationship, and the rate at which brain weight changes with body weight on log-scale is the same for primates and non-primates.

## (c)

**Model Setup:** Here we regress brainwei on bodyweig *without* any logarithmic transformation:

$$\text{brainwei} \;=\; \beta_0 \;+\; \beta_1\,\text{bodyweig} \;+\; \varepsilon.$$

```
. regress brainwei bodyweig

      Source |       SS           df       MS      Number of obs   =        28
-------------+----------------------------------   F(1, 26)        =      0.00
       Model |  1372.62473         1  1372.62473   Prob > F        =    0.9785
    Residual |  48113597.9        26  1850522.99   R-squared       =    0.0000
-------------+----------------------------------   Adj R-squared   =   -0.0384
       Total |  48114970.5        27  1782035.94   Root MSE        =    1360.3

------------------------------------------------------------------------------
    brainwei |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    bodyweig |  -.0004326   .0158853    -0.03   0.978    -.0330853    .0322201
       _cons |   576.3724   265.9121     2.17   0.040     29.78228    1122.963
------------------------------------------------------------------------------

. regress brainwei bodyweig if dinosaur==0

      Source |       SS           df       MS      Number of obs   =        25
-------------+----------------------------------   F(1, 23)        =    151.70
       Model |  41094325.4         1  41094325.4   Prob > F        =    0.0000
    Residual |  6230571.04        23  270894.393   R-squared       =    0.8683
-------------+----------------------------------   Adj R-squared   =    0.8626
       Total |  47324896.4        24  1971870.68   Root MSE        =    520.48

------------------------------------------------------------------------------
    brainwei |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    bodyweig |   .9431658   .0765768    12.32   0.000     .7847546    1.101577
       _cons |   191.2226   110.0878     1.74   0.096    -36.51134    418.9565
------------------------------------------------------------------------------
```

Figure 22: Regression Results including and excluding dinasors

**Including Dinosaurs (All Observations).** When we run the regression on the full sample (including Diplodocus, Triceratops, and Brachiosaurus), the slope is nearly zero ($\hat{\beta}_1 \approx -0.00043$, $p \approx 0.98$) and the model has $R^2 = 0.0000$. In other words, the presence of extremely large dinosaur bodies overwhelms the rest of the data, making a straight line fit on the raw scale practically meaningless.

**Excluding Dinosaurs (Reduced Dataset).** Once we drop all dinosaur observations (`if dinosaur==0`), the regression changes dramatically:

- $\hat{\beta}_1 = 0.943$ with $p < 0.0001$, indicating a significant positive relationship between body weight and brain weight on the arithmetic scale.

- $R^2 = 0.8683$, so nearly 87% of the variance in brainwei is now explained by bodyweig.
- The root MSE is about 520.48, which is much smaller relative to the brain weight range than before.

#### Why This Works Better

- **Extreme Body Weights:** Dinosaurs had body weights of thousands or tens of thousands of kilograms, far beyond the rest of the mammals in the dataset. Without a transformation (like a log transform), these extreme values cause the slope estimate to flatten severely when dinosaurs are included, because the regression tries to accommodate both typical mammals and enormously heavy dinosaurs on the same linear scale.
- **Heterogeneity of Species:** Dinosaurs likely follow a different allometric pattern (brain vs. body growth) than modern mammals. Removing them allows a single linear relationship to capture the mammalian data quite well.

## (d)

#### (i) Regression with the Response in Original Units and the Predictor Logged

First, we regress the `brainwei` (original scale) on ln(bodywt). The Stata output is:

```
. regress brainwei logbodywt

      Source |       SS           df       MS      Number of obs   =        28
-------------+----------------------------------   F(1, 26)        =      5.12
       Model |   7910994.7         1   7910994.7   Prob > F        =    0.0323
    Residual |  40203975.8        26  1546306.76   R-squared       =    0.1644
-------------+----------------------------------   Adj R-squared   =    0.1323
       Total |  48114970.5        27  1782035.94   Root MSE        =    1243.5

------------------------------------------------------------------------------
    brainwei |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   logbodywt |   143.5545   63.46717     2.26   0.032     13.09588    274.0131
       _cons |   33.13352   335.4335     0.10   0.922    -656.3599    722.6269
------------------------------------------------------------------------------
```

Figure 23: Regression Results

#### Interpretation:

- The slope $\hat{\beta}_1 \approx 143.55$ implies that a one-unit increase in ln(bodywt) ($\approx$ multiplying body weight by $e \approx 2.72$) is associated with an additive increase of 143.6 grams in brain weight on average, and the relationship is significant $p = 0.032$.
- The $R^2$ is around 0.1644, indicating that only 16% of the variance in brain weight is explained by logged body weight on its own.

#### (ii) Box-Cox Transformation of the Response

We then apply the user-written `boxcox` command:

```
. boxcox brainwei logbodywt
Fitting comparison model
```

```
Iteration 0:   log likelihood = -240.72687
Iteration 1:   log likelihood = -225.64612
Iteration 2:   log likelihood = -215.07312
Iteration 3:   log likelihood = -187.69228
Iteration 4:   log likelihood = -186.98903
Iteration 5:   log likelihood = -186.98889
Iteration 6:   log likelihood = -186.98889

Fitting full model

Iteration 0:   log likelihood = -238.21209
Iteration 1:   log likelihood = -175.11538
Iteration 2:   log likelihood = -174.53035
Iteration 3:   log likelihood = -174.53025
Iteration 4:   log likelihood = -174.53025

                                                  Number of obs   =         28
                                                  LR chi2(1)      =      24.92
Log likelihood = -174.53025                       Prob > chi2     =      0.000

------------------------------------------------------------------------------
    brainwei |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      /theta |   .0092382   .0620415     0.15   0.882    -.1123609    .1308372
------------------------------------------------------------------------------

Estimates of scale-variant parameters
----------------------------
             |      Coef.
-------------+--------------
Notrans      |
   logbodywt |   .5126691
       _cons |   2.610343
-------------+--------------
      /sigma |    1.53683
----------------------------

---------------------------------------------------------
   Test         Restricted     LR statistic      P-value
    H0:       log likelihood       chi2       Prob > chi2
---------------------------------------------------------
theta = -1      -261.65448       174.25           0.000
theta =  0      -174.54137         0.02           0.881
theta =  1      -238.21209       127.36           0.000
---------------------------------------------------------
```

- $\hat{\lambda} \approx 0.0092$ is very close to 0, and we cannot reject $\lambda = 0$.
- This suggests that ln(brainwei) is an appropriate transformation of the response.
- Other tests ($\lambda = -1$ or $\lambda = 1$) yield significant differences and far worse fit.

**Generating the Box-Cox Transform and Regressing:**

```
. gen BC_brain = (brainwei^0.0092 - 1) / 0.0092

. regress BC_brain logbodywt

      Source |       SS           df       MS      Number of obs   =        28
-------------+----------------------------------   F(1, 26)        =     39.67
       Model |  100.867787         1  100.867787   Prob > F        =    0.0000
    Residual |  66.1093691        26  2.54266804   R-squared       =    0.6041
-------------+----------------------------------   Adj R-squared   =    0.5889
       Total |  166.977156        27  6.18433912   Root MSE        =    1.5946

------------------------------------------------------------------------------
    BC_brain |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   logbodywt |   .5125987   .0813853     6.30   0.000     .3453088    .6798887
       _cons |    2.61011   .4301336     6.07   0.000     1.725958    3.494263
------------------------------------------------------------------------------
```

Figure 24: Regression Results after Box-Cox Transform

Now the slope is $\approx 0.513$ for the Box-Cox–transformed response, with $R^2 \approx 0.60$. This represents a substantial improvement over the partial-log model (which had $R^2 \approx 0.16$). Moreover, because $\theta \approx 0$ from the Box-Cox procedure, it endorses a log transform of brainweight.

### Conclusion and Comparison to the Fully Log–Log Model

- **Box-Cox Conclusion:**
  - The best-fitting $\theta$ is very close to 0, so ln(brainweight) is recommended. This confirms that taking logs of both the response and the predictor (body weight) is the most appropriate transformation.
  - In other words, an additive model on the Box-Cox scale is virtually identical to the standard log–log model.
  - **Interpretation of the Slope:** Since $\theta \approx 0$, the slope coefficient (about 0.51) on log(bodywt) indicates that a 1% increase in body weight is associated with roughly a 0.51% increase in brain weight. In other words, brain weight scales as bodyweight$^{0.51}$.
- **Comparison to Class Results (Log–Log):**
  - In class, using ln(brainweight) vs. ln(bodyweight), the slope was $\approx 0.496$, with $R^2 \approx 0.608$.
  - Here, after the Box-Cox transform, the slope is $\approx 0.51$, and $R^2 \approx 0.60$. These are very similar, indicating a consistent allometric relationship.
  - Hence, both approaches yield nearly the same conclusion: brain weight scales roughly as bodyweight$^{0.5}$. Box-Cox merely validates that ln(brainweight) is the correct form for modeling this dataset.