22401 HW1

Bin Yu

January 21, 2025

Question 1

(a)

To test $H_0 : \beta_1 = 5$ versus $H_1 : \beta_1 \neq 5$ at $\alpha = 0.10$. From the table, the estimated slope is $\hat{\beta}_1 = 7.23$ with standard error $\mathrm{SE}(\hat{\beta}_1) = 0.85$. The test statistic is

$$T = \frac{7.23 - 5}{0.85} \approx 2.62.$$

Compare $|T| = 2.62$ with the critical value $t_{0.05,18} \approx 1.73$ (two-sided, $\alpha = 0.10$). Since $2.62 > 1.73$, **reject** H_0 at the $\alpha = 0.10$ significance level.

(b)

To test $H_0 : \beta_0 = -150$ versus $H_1 : \beta_0 \neq -150$ at $\alpha = 0.10$. The estimated intercept is $\hat{\beta}_0 = -188.49$ with standard error $\mathrm{SE}(\hat{\beta}_0) = 33.32$. The test statistic is

$$T = \frac{-188.49 - (-150)}{33.32} = \frac{-38.49}{33.32} \approx -1.15.$$

Comparing $|T| = |-1.15| = 1.15$ to $t_{0.05,18} \approx 1.73$ (two-sided, $\alpha = 0.10$), we see $|-1.15| < 1.73$, so **fail to reject** H_0.

(c)

To construct a 99% confidence interval for β_1. Using $\hat{\beta}_1 = 7.23$ and $\mathrm{SE}(\hat{\beta}_1) = 0.85$, a 99% CI is given by

$$\hat{\beta}_1 \pm t_{0.005,\,18}\,\mathrm{SE}(\hat{\beta}_1).$$

The critical value $t_{0.005,18} \approx 2.88$. Thus,

$$99\%\text{ CI: } 7.23 \pm 2.88 \times 0.85 \approx [\,7.23 - 2.45,\ 7.23 + 2.45\,] \approx [\,4.78,\ 9.68\,].$$

The 99% CI for β_1 is approximately

$$[4.78,\ 9.68]$$

Why is the confidence interval more useful?
A hypothesis test that focuses on one particular value of β_1 (say, testing $H_0 : \beta_1 = 5$) tells only whether you can reject or fail to reject that specific claim. By contrast, a confidence interval provides a range of plausible values for β_1 at a given confidence level. This is useful for several reasons:

- **Checking if 0 is included:** The CI makes it easy to see whether 0 is among the plausible values for β_1. If the entire interval lies above 0, that indicates a positive and statistically significant slope. If it lies below 0, that indicates a negative and significant slope. If the interval crosses 0, there is insufficient evidence of linear relationship.
- **More insight than a single hypothesis test:** Instead of testing one hypothesis (e.g., $\beta_1 = 5$) at a time, the CI implicitly conducts an infinite set of hypothesis tests across all potential slope values. Any value outside the interval would be rejected at the chosen significance level.
- **Magnitude and precision:** The width of the CI conveys how precise the estimate is (narrow intervals suggest high precision; wide intervals suggest low precision). Moreover, seeing the range of plausible values helps assess whether β_1 could be large enough to be of practical importance or small enough to be negligible.

Question 2

(a)

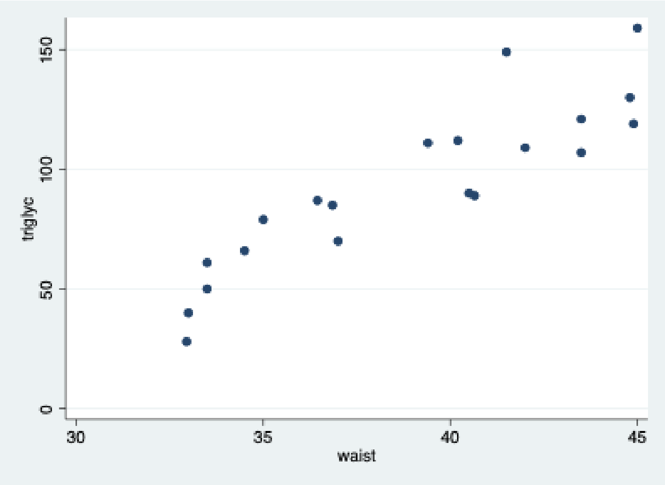
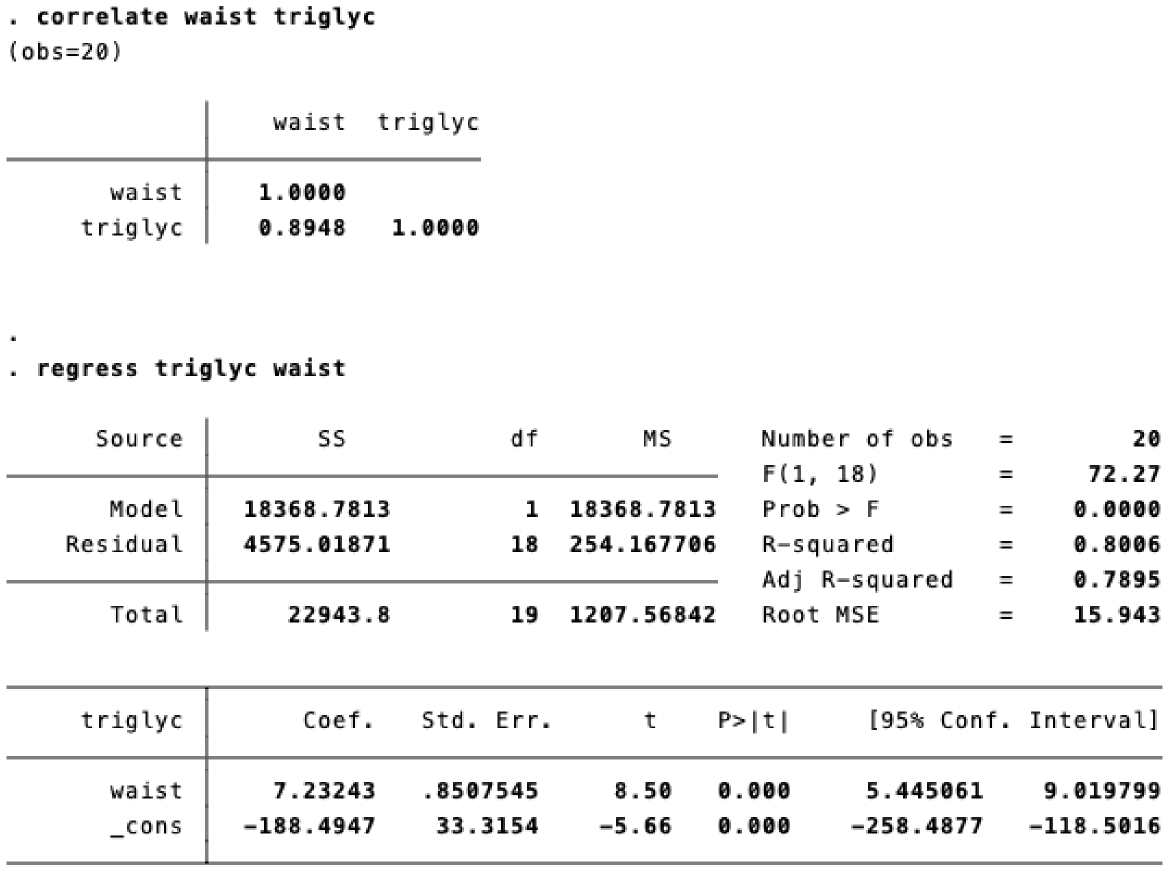


Figure 1: Scatter Plot



```
. correlate waist triglyc
(obs=20)

             |    waist  triglyc
-------------+------------------
       waist |   1.0000
     triglyc |   0.8948   1.0000


.
. regress triglyc waist

      Source |       SS           df       MS      Number of obs   =        20
-------------+----------------------------------   F(1, 18)        =     72.27
       Model |  18368.7813         1  18368.7813   Prob > F        =    0.0000
    Residual |  4575.01871        18  254.167706   R-squared       =    0.8006
-------------+----------------------------------   Adj R-squared   =    0.7895
       Total |     22943.8        19  1207.56842   Root MSE        =    15.943

------------------------------------------------------------------------------
     triglyc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       waist |    7.23243   .8507545     8.50   0.000     5.445061    9.019799
       _cons |  -188.4947    33.3154    -5.66   0.000    -258.4877   -118.5016
------------------------------------------------------------------------------
```

Figure 2: Correlation and Regression

A scatter plot of the two variables `waist` and `triglyc` shows a positive, roughly linear relationship. The sample correlation coefficient (from the Stata output) is

$$r = 0.8948.$$

Regression model.

Here regress `triglyc` on `waist`:

$$\widehat{\texttt{triglyc}} \;=\; \hat\beta_0 + \hat\beta_1\,(\texttt{waist}).$$

From the output:

$$\hat\beta_1 \;=\; 7.2324, \quad \hat\beta_0 \;=\; -188.4947.$$

Relationship between coefficient and slope

Consider a simple linear regression of `triglyc` (denote it by Y) on `waist` (denote it by X):

$$Y = \beta_0 + \beta_1\,X + \varepsilon.$$

In ordinary least squares, the estimated slope $\hat\beta_1$ is given by

$$\hat\beta_1 \;=\; \frac{\mathrm{Cov}(X,\,Y)}{\mathrm{Var}(X)} \;=\; r_{X,Y}\cdot\frac{s_Y}{s_X},$$

where

- $\mathrm{Cov}(X,\,Y)$ is the sample covariance,

- $\text{Var}(X)$ is the sample variance of X,
- $r_{X,Y}$ is the sample correlation between X and Y,
- s_X and s_Y are the sample standard deviations of X and Y, respectively.

two qualitative points:

1. **Sign:** Because s_X and s_Y are always positive, the sign of $\hat{\beta}_1$ directly matches the sign of the sample correlation $r_{X,Y}$.
   - If $r_{X,Y} > 0$, then $\hat{\beta}_1 > 0$, indicating a positive slope: as X increases, Y also tends to increase.
   - If $r_{X,Y} < 0$, then $\hat{\beta}_1 < 0$, indicating a negative slope: as X increases, Y tends to decrease.
2. **Magnitude (Proportionality):** The slope $\hat{\beta}_1$ is proportional to $r_{X,Y}$, with proportionality constant $\frac{s_Y}{s_X}$. This means that for a fixed ratio $\frac{s_Y}{s_X}$, increasing $|r_{X,Y}|$ increases $|\hat{\beta}_1|$. In practical terms:

   $$|\hat{\beta}_1| \;=\; |r_{X,Y}| \cdot \frac{s_Y}{s_X}.$$

   Hence, a stronger correlation (in absolute value) implies a steeper slope, once we account for the relative scales of X and Y.

(b)

We test

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0.$$

The null hypothesis states that there is *no linear relationship* between waist circumference and triglyceride level. The output shows:

$$t = 8.50, \quad P\text{-value} = 0.000.$$

Because the P-value is much smaller than any typical significance level (e.g. $\alpha = 0.05$), we **reject** H_0. Hence, we conclude that $\beta_1 \neq 0$ and that there is a significant positive linear relationship.

How to test using the computer.
In `Stata`, use the following command:

```
. regress triglyc waist
```

Then examine the coefficient of `waist` in the output, its `t`-statistic, and the corresponding `P>|t|` value. If the p-value is sufficiently small (e.g. below 0.05 for a 5%-level test), we reject the null hypothesis $H_0 : \beta_1 = 0$ in favor of the alternative that $\beta_1 \neq 0$. In other words, we conclude there is a statistically significant linear relationship between `waist` and `triglyc`.

(c)

Predicted values and residuals.
From the fitted model,

$$\hat{y}_i \;=\; \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

we can generate $\hat{y}_i$ for each observation i, and then compute the residuals $e_i = y_i - \hat{y}_i$.

Residual plot

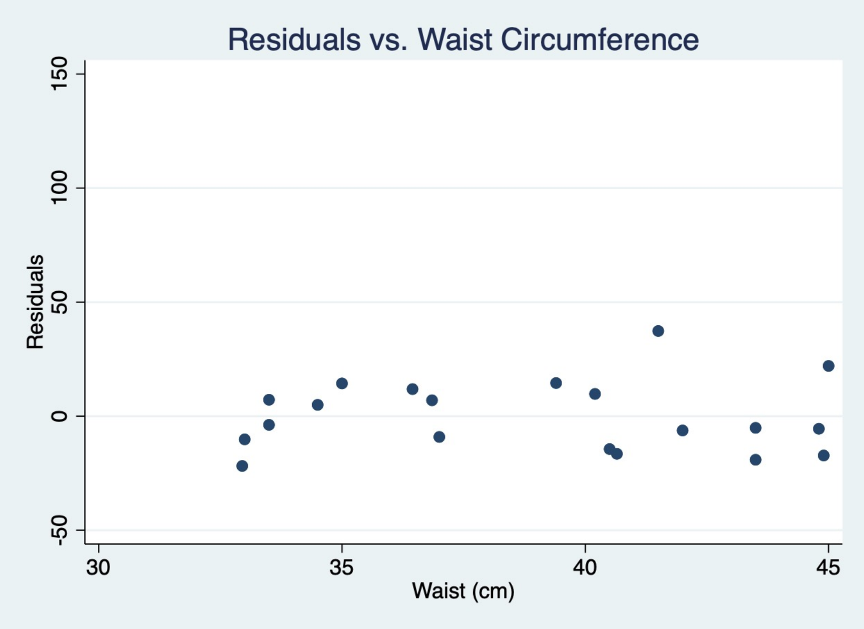


Figure 3: Plot of residuals vs. waist circumference.

If we plot e_i (the residuals) against x_i (`waist`), we expect to see a random scatter around zero, with no obvious pattern. Properties of a good residual plot include:

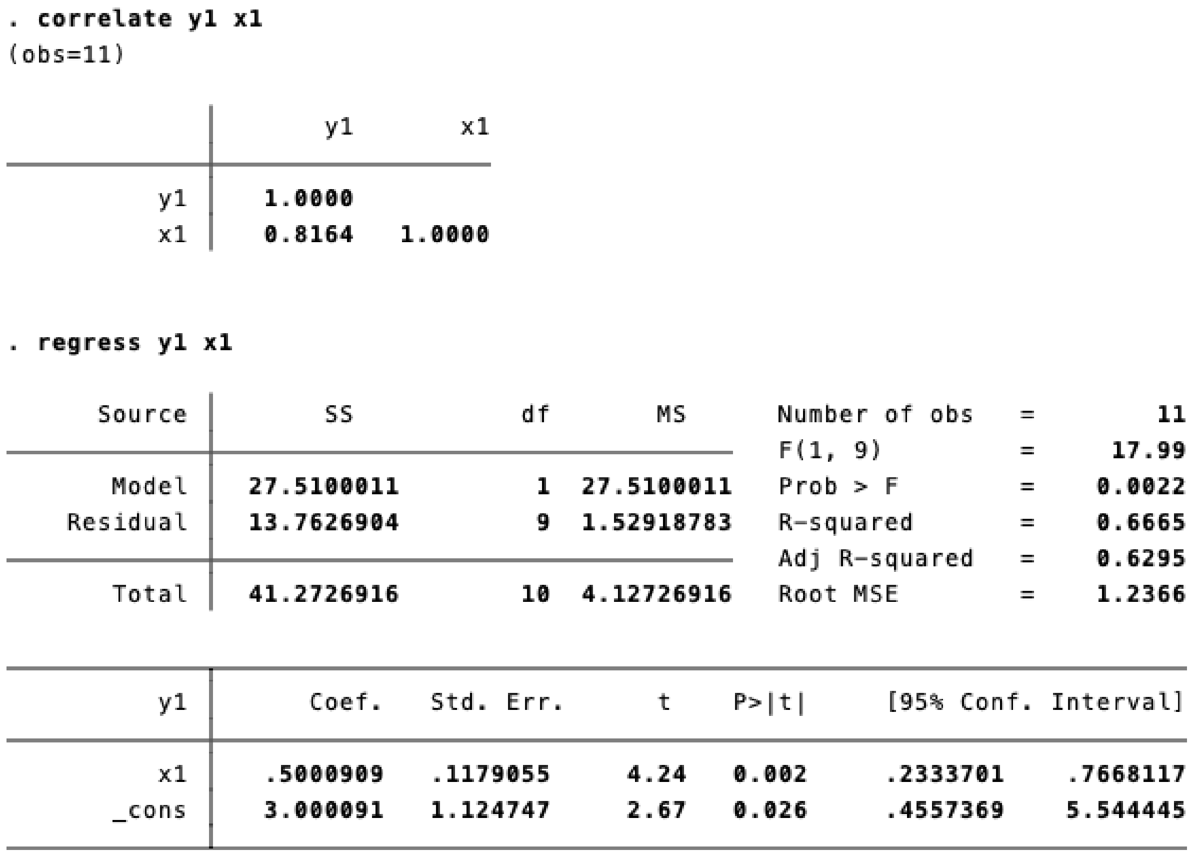
- No clear trend (i.e., no systematic curvature).
- Roughly constant variance of the residuals across the range of x.
- A symmetric distribution of residuals around 0.

From the model's assumptions (linearity, constant variance, etc.), the residual plot now looks reasonably like "white noise" around 0.

If the plot of residuals versus waist circumference shows no clear pattern or trends, then the linear regression model is likely appropriate for these data.

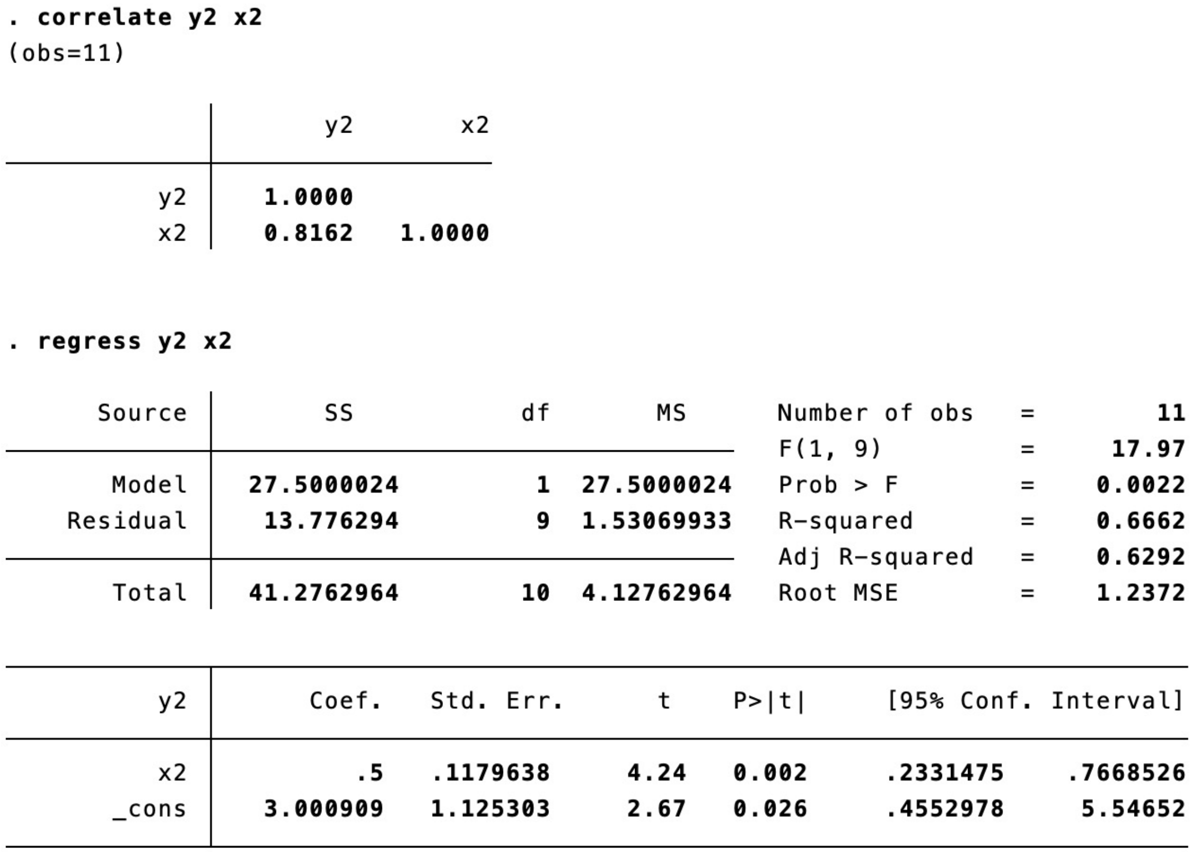
Question 3

(a)



```
. correlate y1 x1
(obs=11)

             |       y1       x1
-------------+------------------
          y1 |   1.0000
          x1 |   0.8164   1.0000


. regress y1 x1

      Source |       SS           df       MS      Number of obs   =        11
-------------+----------------------------------   F(1, 9)         =     17.99
       Model |  27.5100011         1  27.5100011   Prob > F        =    0.0022
    Residual |  13.7626904         9  1.52918783   R-squared       =    0.6665
-------------+----------------------------------   Adj R-squared   =    0.6295
       Total |  41.2726916        10  4.12726916   Root MSE        =    1.2366

------------------------------------------------------------------------------
          y1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .5000909   .1179055     4.24   0.002     .2333701    .7668117
       _cons |   3.000091   1.124747     2.67   0.026     .4557369    5.544445
------------------------------------------------------------------------------
```

Figure 4: y1 to x1



```
. correlate y2 x2
(obs=11)

             |       y2       x2
-------------+------------------
          y2 |   1.0000
          x2 |   0.8162   1.0000


. regress y2 x2

      Source |       SS           df       MS      Number of obs   =        11
-------------+----------------------------------   F(1, 9)         =     17.97
       Model |  27.5000024         1  27.5000024   Prob > F        =    0.0022
    Residual |   13.776294         9  1.53069933   R-squared       =    0.6662
-------------+----------------------------------   Adj R-squared   =    0.6292
       Total |  41.2762964        10  4.12762964   Root MSE        =    1.2372

------------------------------------------------------------------------------
          y2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |         .5   .1179638     4.24   0.002     .2331475    .7668526
       _cons |   3.000909   1.125303     2.67   0.026     .4552978     5.54652
------------------------------------------------------------------------------
```

Figure 5: y2 to x2

I selected two datasets from `anscombe.txt`: the first dataset (x_1, y_1) and another dataset (x_2, y_2). Table 1 summarizes the results, including the intercept, slope, correlation coefficient r, and R^2. These two datasets yield almost identical regression estimates.

Table 1: Regression summary for $y_1 \sim x_1$ and $y_2 \sim x_2$.

	$y_1 \sim x_1$	$y_2 \sim x_2$
Number of observations	11	11
Correlation (r)	0.8164	0.8162
Slope ($\hat{\beta}_1$)	0.5000	0.5000
Intercept ($\hat{\beta}_0$)	3.0009	3.0009
R^2	0.6665	0.6662

(b)

Figures 6 and 7 show the residual plots for the first dataset ($y_1 \sim x_1$). Specifically, Figure 6 is a scatter plot of residuals versus fitted values, and Figure 7 illustrates the kernel density estimate of the residuals (with a normal density overlaid).

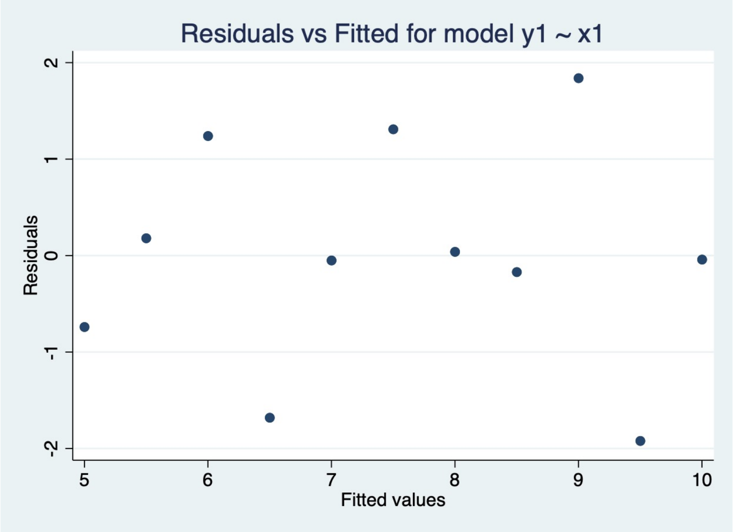


Figure 6: Residuals vs. Fitted for $y_1 \sim x_1$.

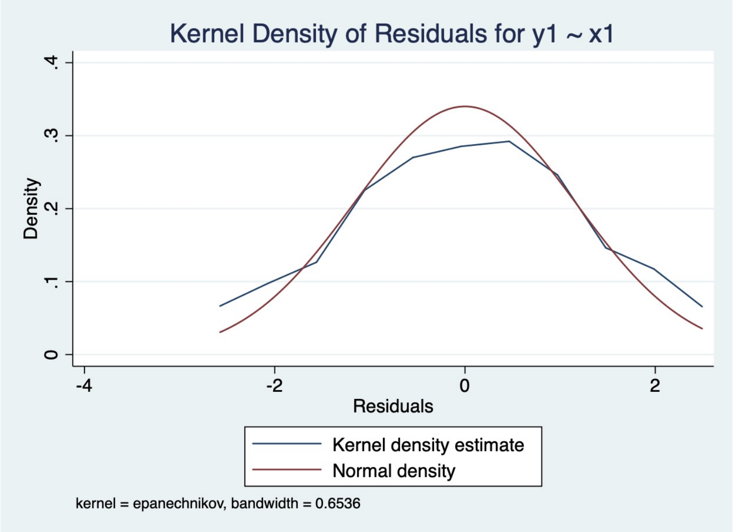


Figure 7: Kernel Density of Residuals for $y_1 \sim x_1$.

From Figure 6, the residuals roughly center around zero without an obvious pattern. In the first dataset, the residuals tend to cluster around zero without obvious systematic structure. "White noise" implies that, on average, the residuals have mean zero and do not exhibit correlations across the fitted values, suggesting that once accounting for the linear relationship, there is no further predictable trend.

The kernel density estimate (Figure 7) suggests that the distribution is slightly skewed but generally normally distributed around 0.

For the second dataset ($y_2 \sim x_2$), the corresponding plots are shown in Figures 8 and 9.

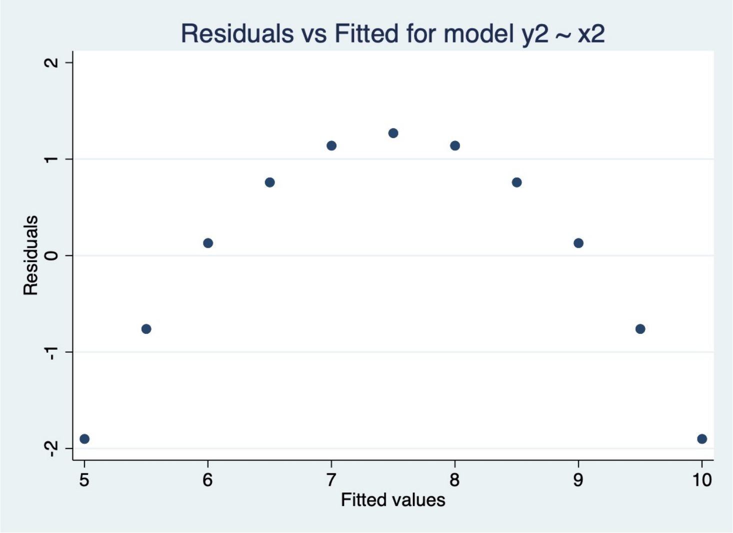


Figure 8: Residuals vs. Fitted for $y_2 \sim x_2$.

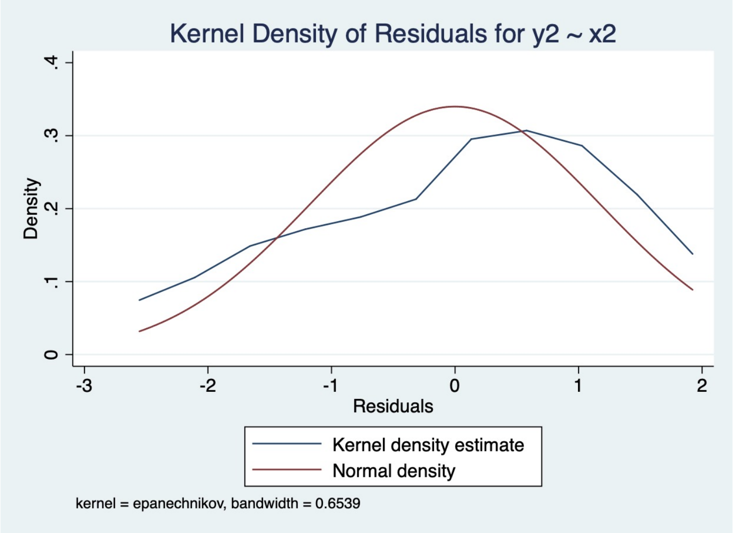


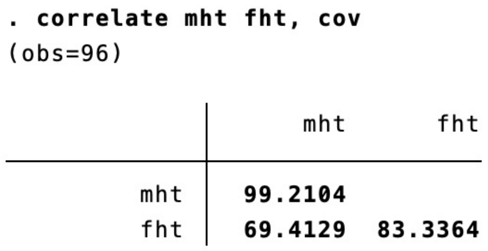
Figure 9: Kernel Density of Residuals for $y_2 \sim x_2$.

Here, even though the slope, intercept, correlation, and R^2 almost match those of the first dataset, the points show a noticeable curve (an arc-like shape) in Figure 8. This indicates that the residuals are not purely random around zero. Instead, they seem to vary systematically with the fitted values.

And the density estimate in Figure 9 shows a slightly different shape compared to a normal distribution. Ideally, residuals should be unpredictable (or normally distributed) if the model fully captures the data's trend. Here, the plot of residuals shows the linear model may be missing some non-linear component (or another underlying structure).

Question 4

(a)



```
. correlate mht fht, cov
(obs=96)

             |      mht      fht
-------------+------------------
         mht |  99.2104
         fht |  69.4129  83.3364
```

Figure 10: Covariance

The covariance between the heights of the husbands (males in this data) and wives (femaies in this data) is 69.4129.

(b)

Suppose originally we measure heights in centimeters and have variables

$$X_{\text{cm}}, \quad Y_{\text{cm}},$$

with sample covariance

$$\text{Cov}\big(X_{\text{cm}}, Y_{\text{cm}}\big).$$

If we convert these to inches via the relation $1\,\text{inch} = 2.54\,\text{cm}$, then

$$X_{\text{in}} \;=\; \frac{X_{\text{cm}}}{2.54}, \quad Y_{\text{in}} \;=\; \frac{Y_{\text{cm}}}{2.54}.$$

Covariance scales by the product of the constants, so

$$\text{Cov}(X_{\text{in}}, Y_{\text{in}}) \;=\; \text{Cov}\Big(\tfrac{X_{\text{cm}}}{2.54}, \tfrac{Y_{\text{cm}}}{2.54}\Big) \;=\; \frac{1}{(2.54)^2}\,\text{Cov}(X_{\text{cm}}, Y_{\text{cm}}).$$

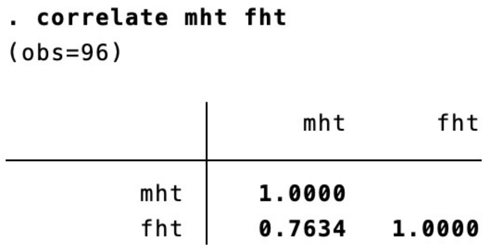
Hence, the covariance in inches is

$$\frac{1}{(2.54)^2} \approx 0.154$$

times the covariance in centimeters.

(c)

]



```
. correlate mht fht
(obs=96)

             |      mht      fht
-------------+------------------
         mht |   1.0000
         fht |   0.7634   1.0000
```

Figure 11: Correlation

The correlation coefficient between the heights of the husbands and wives is 0.7634.

(d)

For any positive constants a and b,

$$\operatorname{corr}(aX, bY) = \frac{\operatorname{Cov}(aX, bY)}{\sqrt{\operatorname{Var}(aX)}\,\sqrt{\operatorname{Var}(bY)}} = \frac{ab\operatorname{Cov}(X, Y)}{|a|\sqrt{\operatorname{Var}(X)}\,|b|\sqrt{\operatorname{Var}(Y)}} = \frac{ab}{ab}\,\frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}} = \operatorname{corr}(X, Y).$$

Thus converting centimeters to inches does not change the correlation.

(e)

Consider a hypothetical scenario where each woman is exactly 5 cm shorter than her husband. Mathematically, let

$$Y \;=\; X - 5.$$

Then

$$\operatorname{Cov}(X, Y) \;=\; \operatorname{Cov}(X,\, X - 5) = \operatorname{Cov}(X, X) - \operatorname{Cov}(X, 5).$$

Since $\operatorname{Cov}(X, X) = \operatorname{Var}(X)$ and $\operatorname{Cov}(X, 5) = 0$ (a constant has no covariance),

$$\operatorname{Cov}(X, Y) = \operatorname{Var}(X).$$

Also,

$$\sqrt{\operatorname{Var}(X)} \;=\; \sqrt{\operatorname{Var}(Y)},$$

because $Y = X - 5$ has the same variance as X. Hence the correlation becomes

$$\operatorname{corr}(X,\, X - 5) = \frac{\operatorname{Cov}(X, X - 5)}{\sqrt{\operatorname{Var}(X)}\,\sqrt{\operatorname{Var}(X - 5)}} = \frac{\operatorname{Var}(X)}{\operatorname{Var}(X)} = 1.$$

In other words, if each wife is exactly 5 cm shorter than her husband, the correlation is +1, which is a pure linear relationship.

(f)

I think my primary interest is in predicting how tall a wife is likely to be, given knowledge of her husband's height. One practical argument is that it might be easier to measure or observe men first (e.g., men often come in for a physical exam earlier), and then we want to estimate the wife's height once we have the husband's data. Consequently, we choose the wife's height `(fht)` as the *dependent* (response) variable, and the husband's height `(mht)` as the *independent* (predictor) variable:

$$\text{fht} \;=\; \beta_0 \;+\; \beta_1\,\text{mht} \;+\; \varepsilon,$$

where:

- fht (the wife's height) is the response we want to predict,
- mht (the husband's height) is the explanatory factor, and
- ε is the random error term.

This setup allows us to interpret β_1 as the expected change in a wife's height (in cm) for a one-centimeter increase in the husband's height, holding other factors constant (to the extent possible).

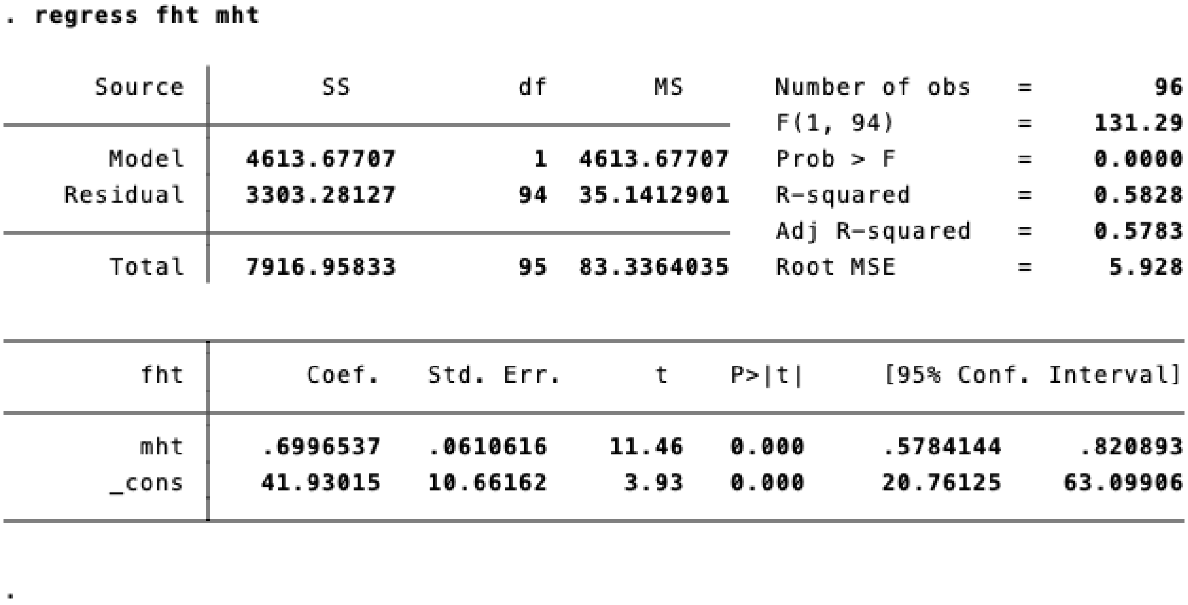
(g)

Assume we regress wife's height (fht) on husband's height (mht):

$$fht \;=\; \beta_0 \;+\; \beta_1\,mht \;+\; \varepsilon.$$

We test

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_a : \beta_1 \neq 0.$$



```
. regress fht mht

      Source |       SS           df       MS      Number of obs   =        96
-------------+----------------------------------   F(1, 94)        =    131.29
       Model |  4613.67707         1  4613.67707   Prob > F        =    0.0000
    Residual |  3303.28127        94  35.1412901   R-squared       =    0.5828
-------------+----------------------------------   Adj R-squared   =    0.5783
       Total |  7916.95833        95  83.3364035   Root MSE        =     5.928

------------------------------------------------------------------------------
         fht |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mht |   .6996537   .0610616    11.46   0.000     .5784144     .820893
       _cons |   41.93015   10.66162     3.93   0.000     20.76125    63.09906
------------------------------------------------------------------------------

.
```

Figure 12: Regression of wife's height (fht) on husband's height (mht).

From the regression output (see Figure 12), the estimated slope is

$$\hat{\beta}_1 \approx 0.70 \quad (\text{t-stat} = 11.46,\; p < 0.0001),$$

and the 95% confidence interval for β_1 is approximately $[0.58,\ 0.82]$. Since the p-value is extremely small (essentially 0.0000), we *reject* H_0 and conclude that there is a statistically significant linear relationship between husbands' and wives' heights. For each additional centimeter in a husband's height, the expected increase in his wife's height is about 0.70 cm, on average.

If we fail to reject H_0, it would mean that there was no sufficient evidence in the sample that there is a linear relationship between husbands' and wives' heights.

(h)

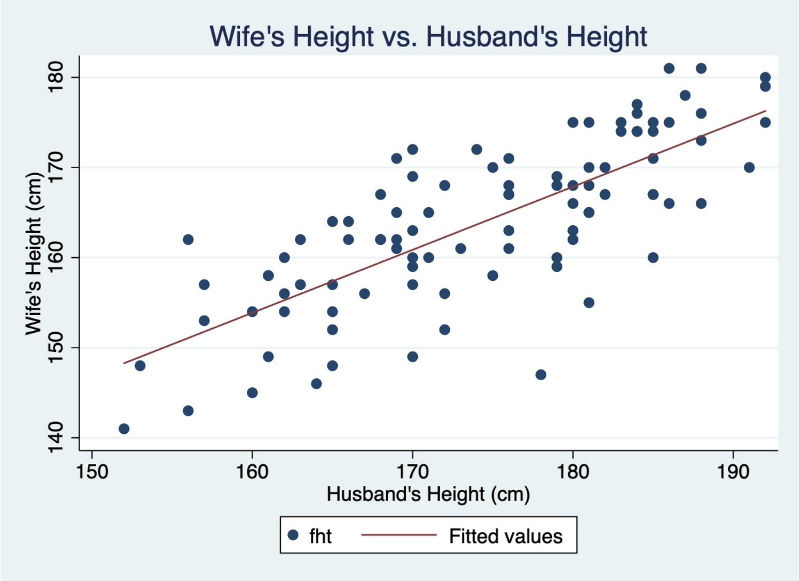


Figure 13: Scatter plot of wife's height vs. husband's height. Each point represents one couple, with the red line indicating the fitted OLS regression.

Comment: Figure 13 shows a positive relationship between husbands' heights (x-axis) and wives' heights (y-axis). As husbands become taller, wives also tend to be taller on average. The majority of points cluster around the regression line without obvious curvature or extreme outliers. This visual evidence supports the conclusion that taller men generally marry taller women (and, conversely, shorter men marry shorter women).

(i)

Model and Hypotheses

We fit a simple linear regression model:

$$\text{fht} = \beta_0 \ + \ \beta_1 \,\text{mht} + \varepsilon,$$

where

- fht = wife's height (in cm),
- mht = husband's height (in cm),
- ε = random error term.

To test whether "taller men marry taller women," we set up a one-sided hypothesis on the slope β_1:

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_a : \beta_1 > 0.$$

Under H_0, there is no linear relationship between husband's and wife's heights. Under H_a, wife's height increases as husband's height increases.

Estimation and Test Statistic

Using ordinary least squares on the sample of $n = 96$ couples, we obtain:

$$\hat{\beta}_1 \;=\; 0.6997, \quad \text{SE}(\hat{\beta}_1) \;=\; 0.0611,$$

with a corresponding t-value of

$$t_{\text{obs}} \;=\; \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)} \;=\; \frac{0.6997}{0.0611} \;\approx\; 11.46.$$

Under the null hypothesis $H_0\colon \beta_1 = 0$, the test statistic follows (approximately) a t-distribution with $(n-2)$ degrees of freedom.

Here, For $df = 94$, the typical critical values for one-sided α are:

$$\begin{aligned} t_{0.05,\,94} &\approx 1.66 \quad (\alpha = 0.05), \\ t_{0.01,\,94} &\approx 2.36 \quad (\alpha = 0.01), \\ t_{0.001,\,94} &\approx 3.16 \quad (\alpha = 0.001), \end{aligned}$$

Our observed statistic is

$$t_{\text{obs}} = 11.46,$$

which is larger than 3.16. Therefore, our p-value is small. Thus we reject H_0 in favor of H_a. Statistically, this implies that taller men do, on average, marry taller women, and shorter men marry shorter women.

Moreover, the estimated slope of about 0.70 suggests that for each additional centimeter of the husband's height, the expected wife's height increases by approximately 0.70 cm, on average.

Question 5

(a)

Because the parasite count varies greatly, need to transform `number` using the natural logarithm.

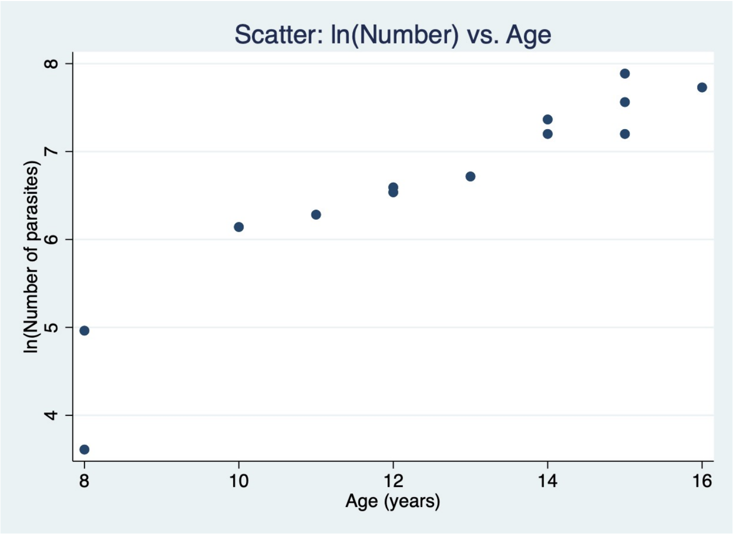
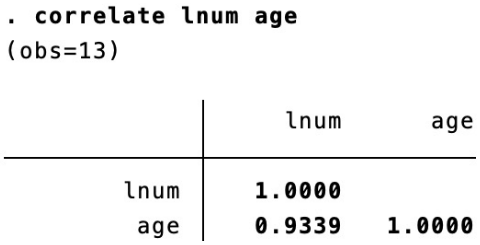


Figure 14: Scatter plot of ln(number) (y-axis) vs. age (x-axis).

Comment: We see a association: as `age` increases, ln(`number`) also tends to increase. Using the natural log helps stabilize the large variability observed on the original scale.

(b)

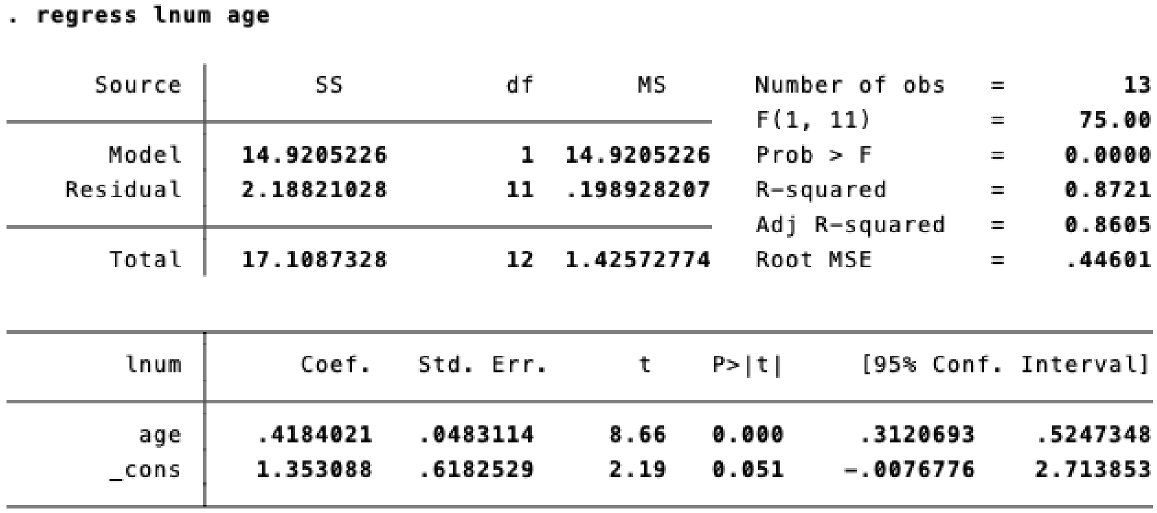


```
. correlate lnum age
(obs=13)

             |     lnum      age
-------------+------------------
        lnum |   1.0000
         age |   0.9339   1.0000
```

Figure 15: Correlation

Comment: A positive correlation would indicate that older children tend to have higher logged parasite counts.

(c)



```
. regress lnum age

      Source |       SS           df       MS      Number of obs   =        13
-------------+----------------------------------   F(1, 11)        =     75.00
       Model |  14.9205226         1  14.9205226   Prob > F        =    0.0000
    Residual |  2.18821028        11  .198928207   R-squared       =    0.8721
-------------+----------------------------------   Adj R-squared   =    0.8605
       Total |  17.1087328        12  1.42572774   Root MSE        =    .44601

------------------------------------------------------------------------------
        lnum |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .4184021   .0483114     8.66   0.000     .3120693    .5247348
       _cons |   1.353088   .6182529     2.19   0.051    -.0076776    2.713853
------------------------------------------------------------------------------
```

Figure 16: Regression

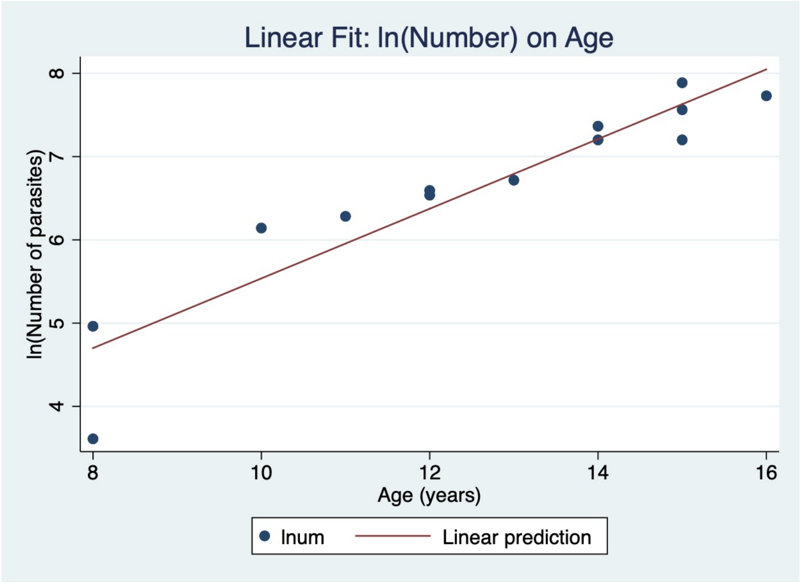


Figure 17: Scatter plot of ln(number) vs. age, with OLS regression line.

Comment: The slope and intercept in this model describe the linear relationship on the log-scale. The slope is significantly positive, it suggests that each one-year increase in age corresponds to an exponential increase in the expected value of `number` of parasites (in the original scale).

(d)

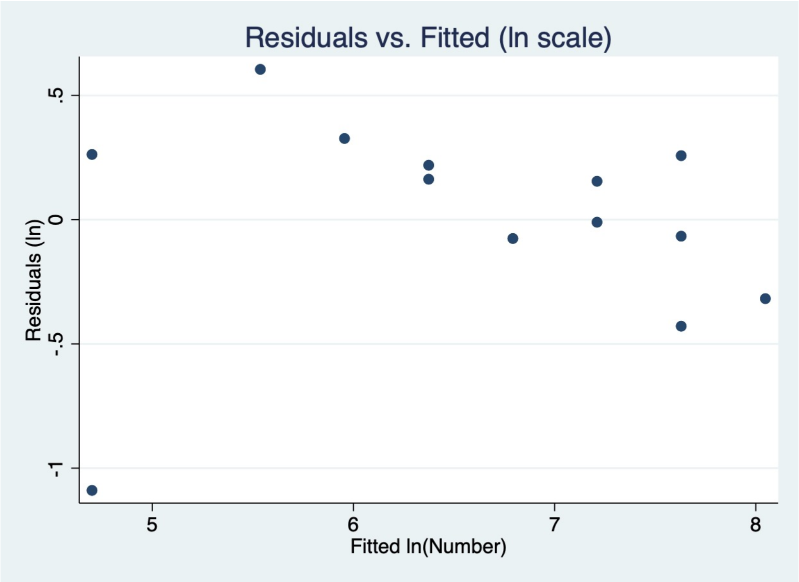


Figure 18: Plot of residuals vs. fitted values on the log scale.

Comment: These residuals scatter randomly around zero, with no obvious pattern, therefore, the model fits well. Any strong curvature, trend, or funnel shape could indicate a model misspecification or non-constant variance.

(e)

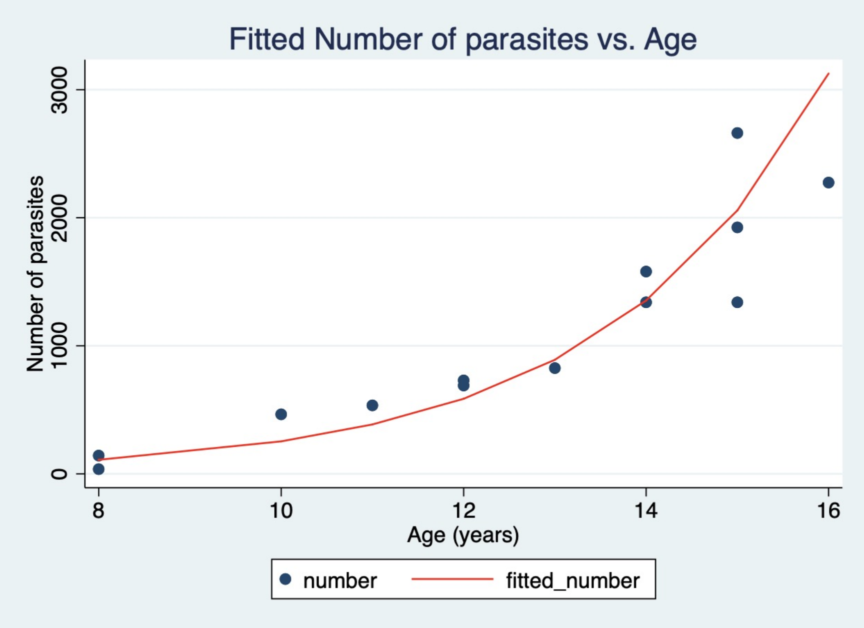


Figure 19: Scatter plot of `number` vs. `age` in the original scale

Comment: The red line shows how many parasites we expect, on average, for each age. By using ln(number) in the regression, we are fitting an exponential-type growth curve in the original scale.