Lecture 7: The Relationship Between Correlation and Regression

C91AR: Advanced Statistics using R

2025-03-04

Contents

1 Reading for today

Barr (2025)

Canduela and Raeside (2020)

2 Session outline

- Today, we are continuing to learn about statistics through data simulation.
- We also continue to use the "heights_and_weights.csv" dataset we used for the session on simulating correlation data.
- We are moving on to look at how you can estimate values given the statistics of simple regression.

3 Regression Theory

- Regression can be thought of as a process of generating a line of best fit given the data.

- Rather than draw by hand, a procedure called **least squares regression** can be used to reliably draw this line.

3.1 Least squares regression

The line is adjusted to minimise the SSE (sum of squared deviations; e_i^2) of each point from the line:

Minimise SSE $= min \sum_{i=1}^{n} e_i^2 = min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

- y_i = observed value
- $\hat{y}_i$ = predicted value
- e_i = error
- n = number of observations
- $\sum_{i=1}^{n}$ = summation of a sequence of terms indexed by i, starting from $i = 1$ and ending at $i = n$

4 Illustration of regression

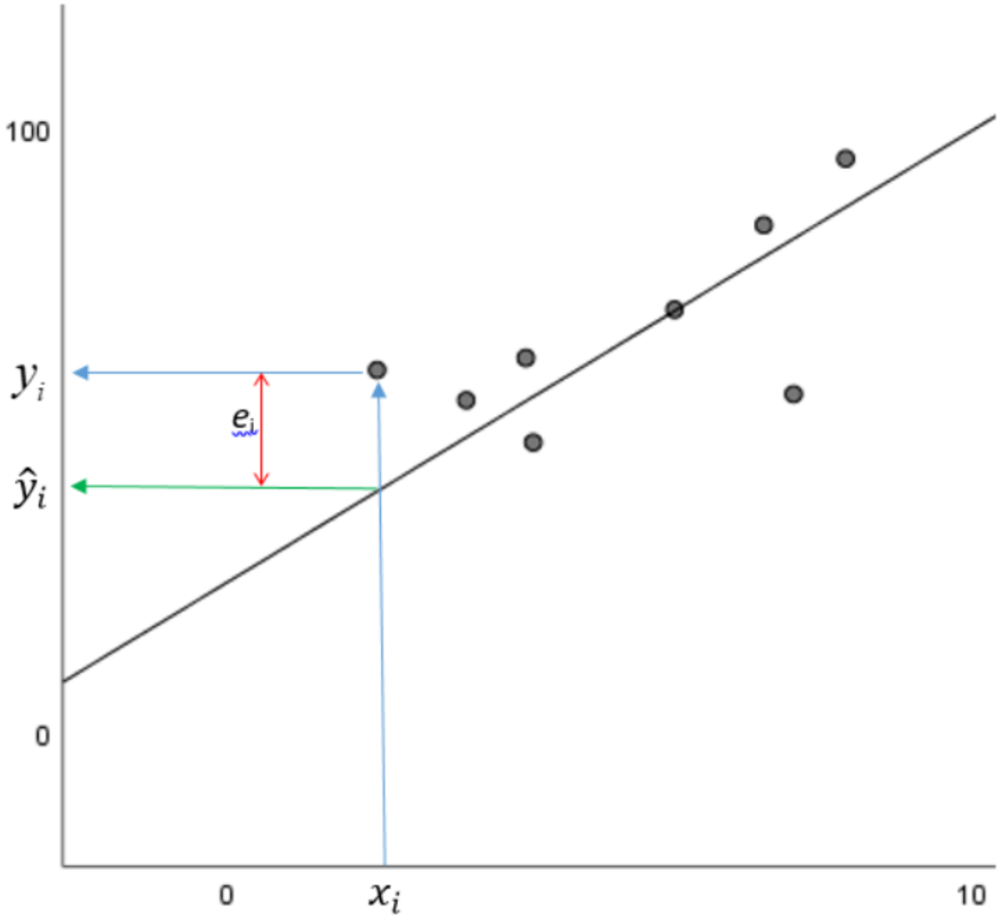


Figure 8.3 Observed, predicted and error values

- So, least squares regression is summation of all of the squared differences (e_i) between the observed (y_i) and predicted ($\hat{y}_i$) values of y

5 Regression formula

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

Where…

- y_i = ith observation of the dependent variable, $i = 1, 2, ..., n$
- x_i = ith observation for the independent variable, $i = 1, 2, ...n$
- β_0 = intercept or constant
- β_1 = the slope or gradient of the predictor/independent variable
- ϵ_i = error or residual of the ith observation
- n = total number of observations

6 The relationship between correlation and regression?

- Correlation tells us about the strength and relationship between two variables.
- Regression (on the other hand) predicts the value of one variable based on the value of another variable.
- For example, in the heights and weights data, we can use regression modelling to *predict someone's weight given their height.*

7 Reproducibility & notation

```r
# Set seed for reproducibility
set.seed(123)

# Change the output format
options(scipen = 999)
```

8 Load packages

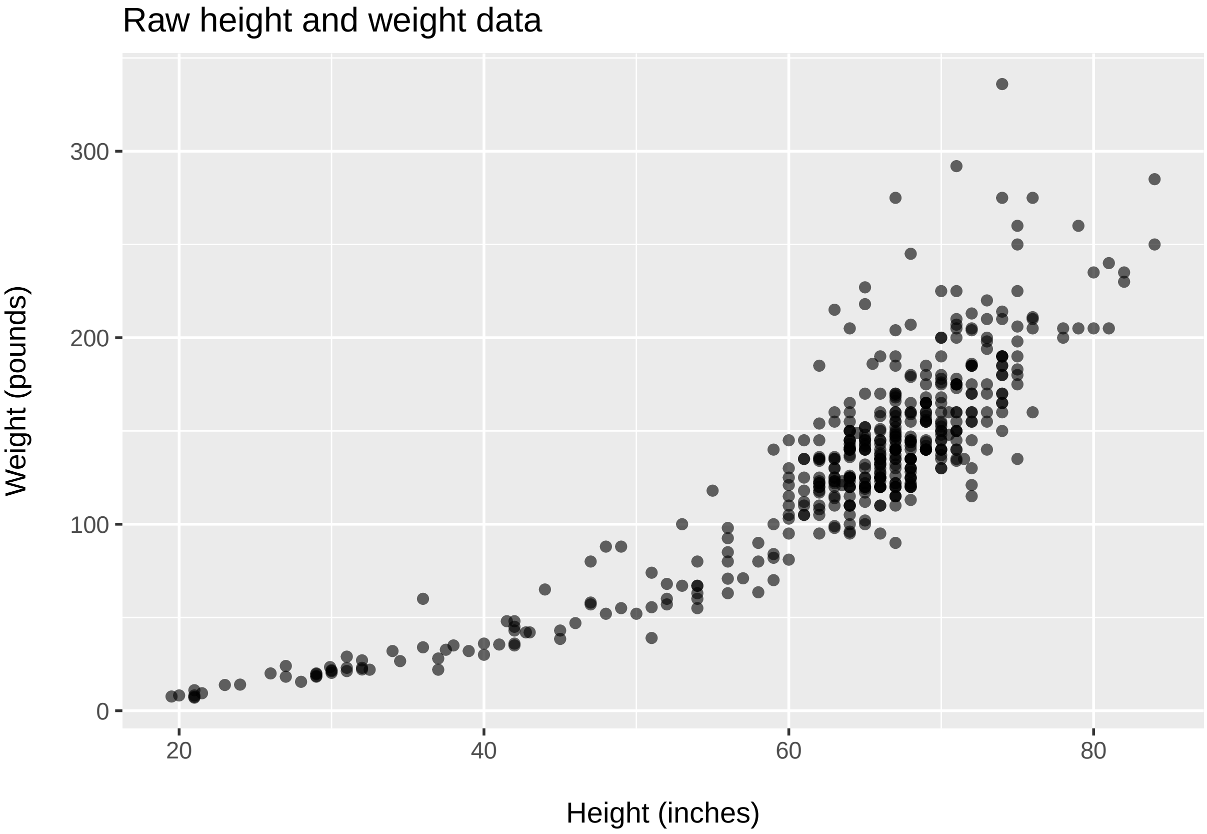
```
# Load packages
pacman::p_load(tidyverse,
               corrr,
               tidyplots)
```

9 Data setup

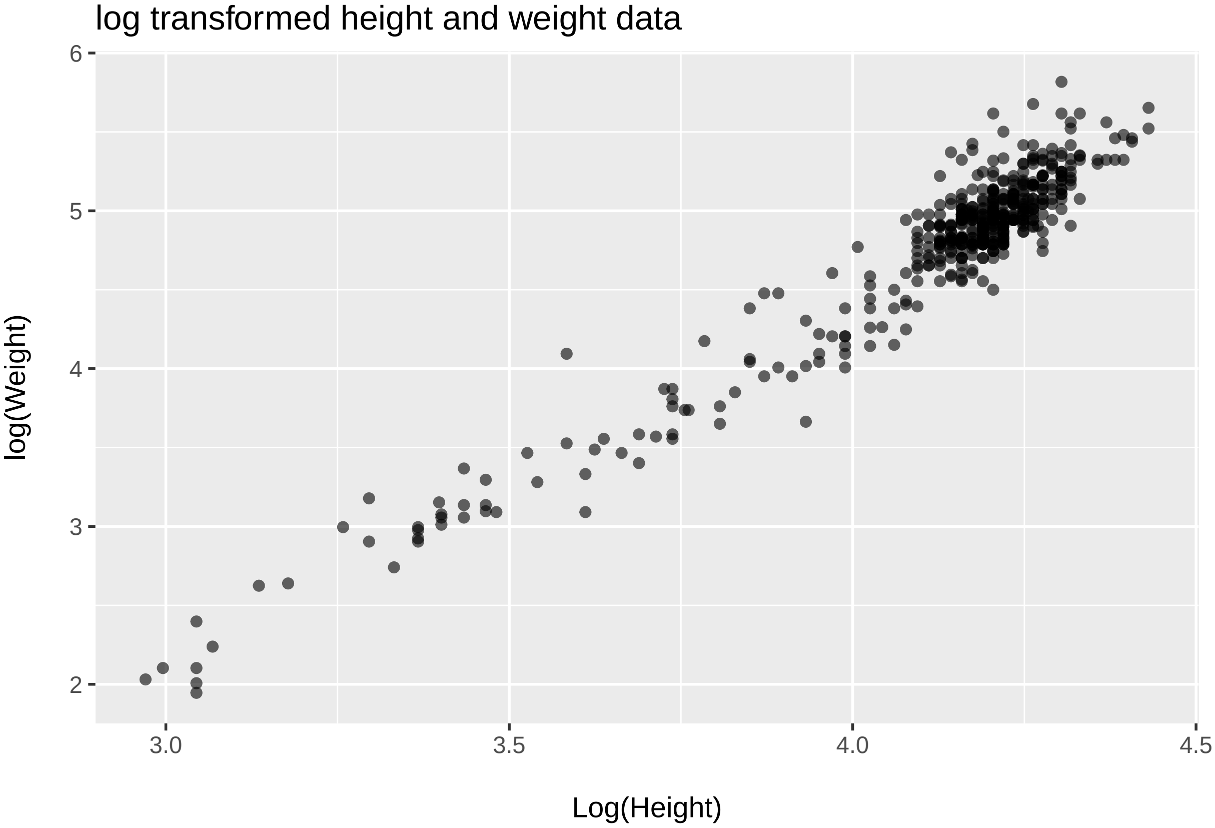
```
# Read in the data
handw <-
  read_csv("data_raw/heights_and_weights.csv",
           col_types = "dd")



# Add log transformed vectors to dataset
handw_log <-
  handw |>
  mutate(hlog = log(height_in),
         wlog = log(weight_lbs))
```

10 Remind me, why are we using the log of the data?

11 Raw data



12 Normalised data



13 Rationale for taking the log

- When we take the log of the data we normalise it, to stabilise the variance in our vectors.
- This is particularly useful if the data are highly skewed or show **heteroscedasticity**
  - When the difference between the observed and predicted values (i.e., the residuals e_i) is not constant across all levels of the independent variables.
  - For the heights and weights data, you can see that the residuals increase as the values for height and weight approach the upper limit
- Normalised data is more suitable for statistical analysis that assume normality, such as linear regression.

14 Using regression to make predictions based on height and weight data

Regression formula

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

In the context of predicting height from weight

- Y_i = prediction of a person i's weight
- X_i = observed height
- β_0 = y-intercept
- β_1 = slope parameter
- e_i = residuals
  - Note, it is assumed that e_i comes from a normal distribution with a mean of zero and variance σ^2.

15 Making predictions using the available statistics

To estimate the parameters of the regression between the y-intercept (β_0) and the slope (β_1) all we need is the:

- Mean estimates for X and Y, denoted as $\hat{\mu_x}$ & $\hat{\mu_y}$
- Standard deviations for X and Y, denoted as $\hat{\sigma_x}$ & $\hat{\sigma_y}$
- Correlations between X and Y, denoted as $\hat{\rho}$
- Note: the $\hat{hat}$ denotes an estimate of a population parameter.

So, the statistics required to estimate β_0 and β_1 are much the same as we used for simulating correlational data.

16 A reminder of our previous calculations

- $\hat{\mu_x} = 4.11, \hat{\sigma_x} = 0.26$ (estimated mean and SD of log height)
- $\hat{\mu_y} = 4.74, \hat{\sigma_y} = 0.65$ (estimated mean and SD of log weight)
- $\hat{\rho_{xy}} = 0.96$ (estimated correlation between the two)

17 Estimating the slope β_1

- Let's start by estimating the value of the slope β_1.
- Importantly, β_1 can be expressed in terms of the correlation coefficient ρ times the ratio of the standard deviations of Y and X.

$$\beta_1 = \rho \frac{\sigma_Y}{\sigma_X}$$

18 Plugging in the numbers

- Now, you can use the estimates of log height and log weight, to estimate the slope:

$$\beta_1 = \rho \frac{\sigma_Y}{\sigma_X}$$

In R Code:

```
# Estimate slope using formula
b1 <- .96 * (.65 / .26)
b1 # 2.4
```

```
## [1] 2.4
```

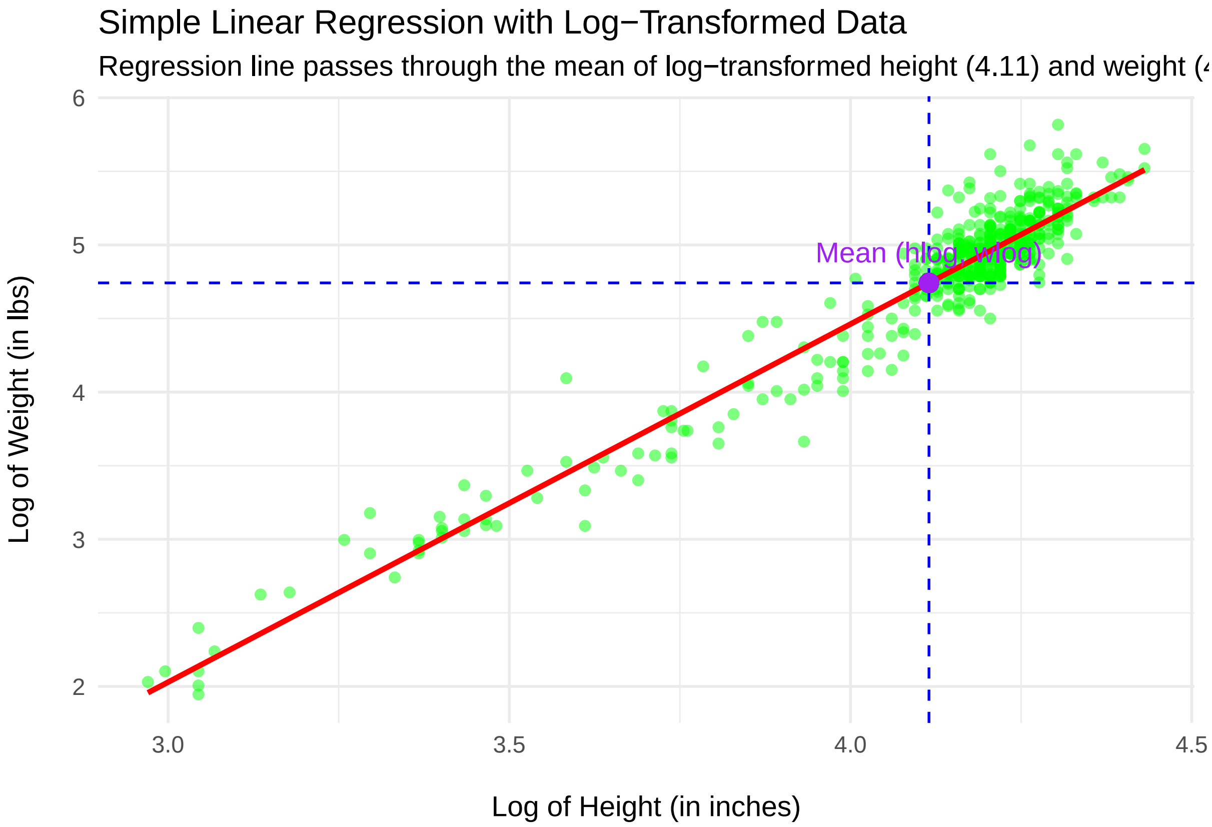
19 Using the Axis to fill in the blanks: part 1

- For mathematical reasons, the regression line is **guaranteed to go through the point corresponding to the mean of both X and Y, i.e., the point (μ_x, μ_y)**.
- One way to think about this is that the regression line pivots around that point depending on the slope (β_1).
- We also know that β_0 is the y-intercept, where the line crosses the vertical axis at $X = 0$.

20 Regression line passing through variable means (μ_x, μ_y)

```
## Warning in geom_point(aes(x = mean_hlog, y = mean_wlog), color = "purple", : All aesthetics have l
## i Please consider using `annotate()` or provide this layer with data containing
##   a single row.
```

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



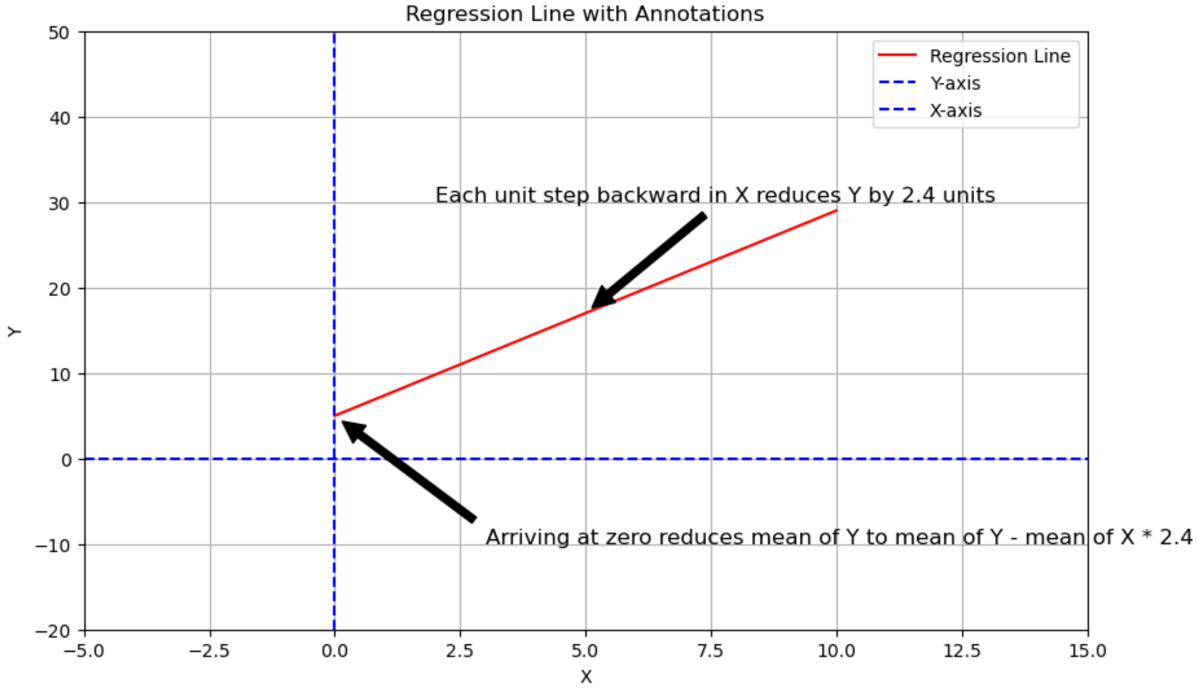
21 Using the Axis to fill in the blanks: part 2

- From all of this information we can calculate β_0.
- Remember that β_1 tells you that for each change in X you have a corresponding change of **2.4** for Y, and that the line goes through points (μ_x, μ_y) as well as the y-intercept $(0, \beta_0)$.

22 Re-framing the calculations

- Think about stepping back unit-by-unit from the mean of $X = \mu_x$ to $X = 0$.
- At $X = \mu_x$, $Y = 4.74$, because as stated earlier, the regression line is guaranteed to go through the point corresponding to the mean of both X and Y, i.e., the point (μ_x, μ_y) or $(4.11, 4.74)$.
- Each unit step you take backward in the X dimension, Y will reduce by $\beta_1 = 2.4$ units.
- When you get to zero, Y will have dropped from μ_y to $\mu_y - \mu_x\beta_1$.

23 Illustrating this relationship



24 The Solution

- With all of the above considerations taking into account the solution is $\beta_0 = \mu_y - \mu_x \beta_1$.
- Using this information we can calculate the slope value: $\beta_0 = 4.74 - 4.11 \times 2.4 = -5.124$
- Now we have the following statistics:
  - $\beta_1 = 2.4$
  - $\mu_x = 4.11$
  - $\mu_y - 4.74$
  - $\beta_0 = -5.124$

25 Plugging in the numbers

So..

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

Becomes...

$$Y_i = -5.124 + 2.4 X_i + e_i$$

26 Checking the results

- To check the results, let’s first run a regression on the log transformed data using `lm()`, which estimates parameters using *ordinary least squares (OLS) regression.*
- In OLS regression, the goal is to find the line that best fits the data by minimizing the sum of the squared differences between the observed values of the dependent variable and the values predicted by the regression line.
- Note, you are interested in the `Estimate` values.

```
summary(lm(wlog ~ hlog,
           data = handw_log))
```

```
##
## Call:
## lm(formula = wlog ~ hlog, data = handw_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63296 -0.09915 -0.01366  0.09285  0.65635
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -5.26977    0.13169  -40.02 <0.0000000000000002 ***
## hlog         2.43304    0.03194   76.17 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1774 on 473 degrees of freedom
## Multiple R-squared:  0.9246, Adjusted R-squared:  0.9245
## F-statistic:  5802 on 1 and 473 DF,  p-value: < 0.00000000000000022
```

27 Matching the regression output to our calculations

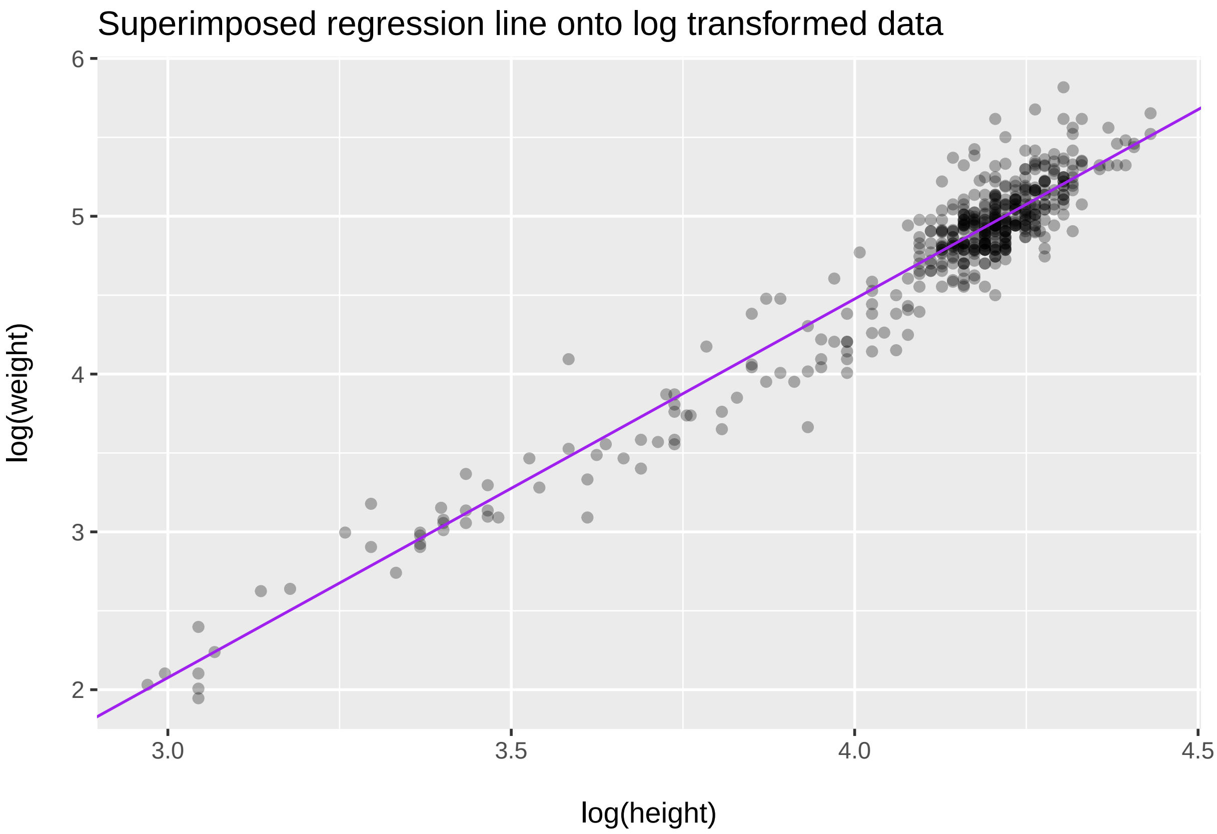
From the model output:

- Estimated slope parameter $\beta_1 = 2.433$ (2.4 from our calculations)
- Estimated y-intercept $\beta_0 = -5.269$ (-5.124 from our calculations)
- These don’t match exactly because of the rounding we’ve used in our calculations.

28 Checking your regression estimate with a plot

Another way to check the accuracy of your regression calculations is to superimpose the regression line on the scatter-plot of the log transformed data.

```
ggplot(data = handw_log,
       aes(hlog, wlog)) +
  geom_point(alpha = .3)+
  geom_abline(intercept = -5.124,
              slope = 2.4,
              colour = 'purple') +
  labs(title = "Superimposed regression line onto log transformed data",
       x = "\nlog(height)",
       y = "log(weight)\n")
```



29 Predicting someone's weight given their height

- Say we want to predict the weight of someone who is 69inches or 175cm (average height of a person from the US). Let's plug the log of this value (4.23) into our regression formula:

$Y_i = -5.124 + 2.4 \times 4.23 + e_i$

- Note: We do not need to provide the residuals (e_i) as they are estimated from the regression equation.

$Y_i = 5.028$

$exp(5.028) = 152.63lbs = 69.2kg$

- So, our regression model predicts that someone who is 175cm tall would weigh 69.2kg.

30 A little more about e_i

- Conventionally, e_i come from a normal distribution with $\mu = 0$ and variance σ^2.
- e_i are important for assessing the model's performance and diagnostic purposes but they are not necessary for making predictions using the regression equation.

31 Model fit

- We did not cover how to test the model's fit today
- This is covered in the RM3 textbook, and there are good resources elsewhere
- I prioritised familiarity with the modelling process over assumption checking

32 Roundup

- Today I have shown you how to calculate a simple regression model by hand using formula
- The aim was to unveil some of the computation that goes on behind the scene to help you understand what regression analysis is actually doing.
- Using this method, we also made a prediction about someone's height given their weight.
- On Monday, we are going to look at multiple regression

33 R^2 Coefficient of determination

To this point we have created a regression equation and used it to predict someone's weight given their height. But, we also want to know how good a fit our equation is given the data. To calculate this we use the ***coefficient of determination*** (R^2):

$$R^2 = \frac{\text{Sum of Squares Explained by Regression (SSR)}}{\text{Total Sum of Squares (before regression)(TSS)}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

The total variation (TSS) in the dependent variable (weight) is split into two parts: the part explained (SSR) by the association between the dependent (weight) and independent (height) variables, and the part that is unexplained (SSE) since the relationship is never perfect and there are always some residuals.

34 R^2 Thresholds

The higher the R^2 value (at least 70%) the better the model fit. A reasonable model fit would be more $R^2 >= 60\%$.

The coefficient of determination can take values between 0 and 1, but is commonly reported as a percentage, as it represent the proportion of the variation in the dependent variable (Y_i) which is explained by the predictor/independent variable (X_i).

35 Exploring the Errors

To diagnose the quality of the model further we need to look at the errors or residuals $(y_i - \hat{y}_i)$ after running the regression model.

Model residuals should be **randomly scattered** with **no extreme values** and should have a **mean of zero**.

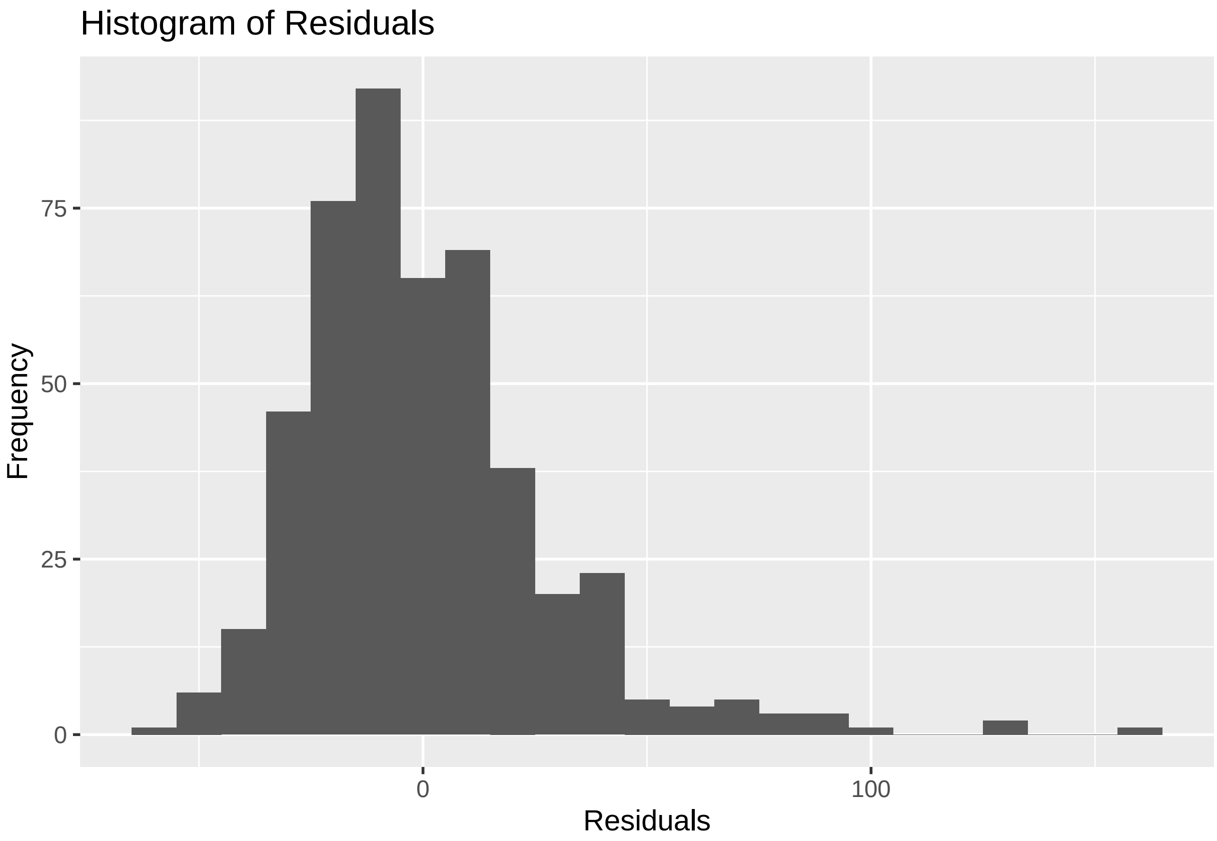
Should these requirements not be met we would have to further investigate whether there is information in the residuals that could be covered by the model or be considered a cause for concern.

A histogram of the residuals and normal probability plot can help you decide how well your residuals fit into the model.

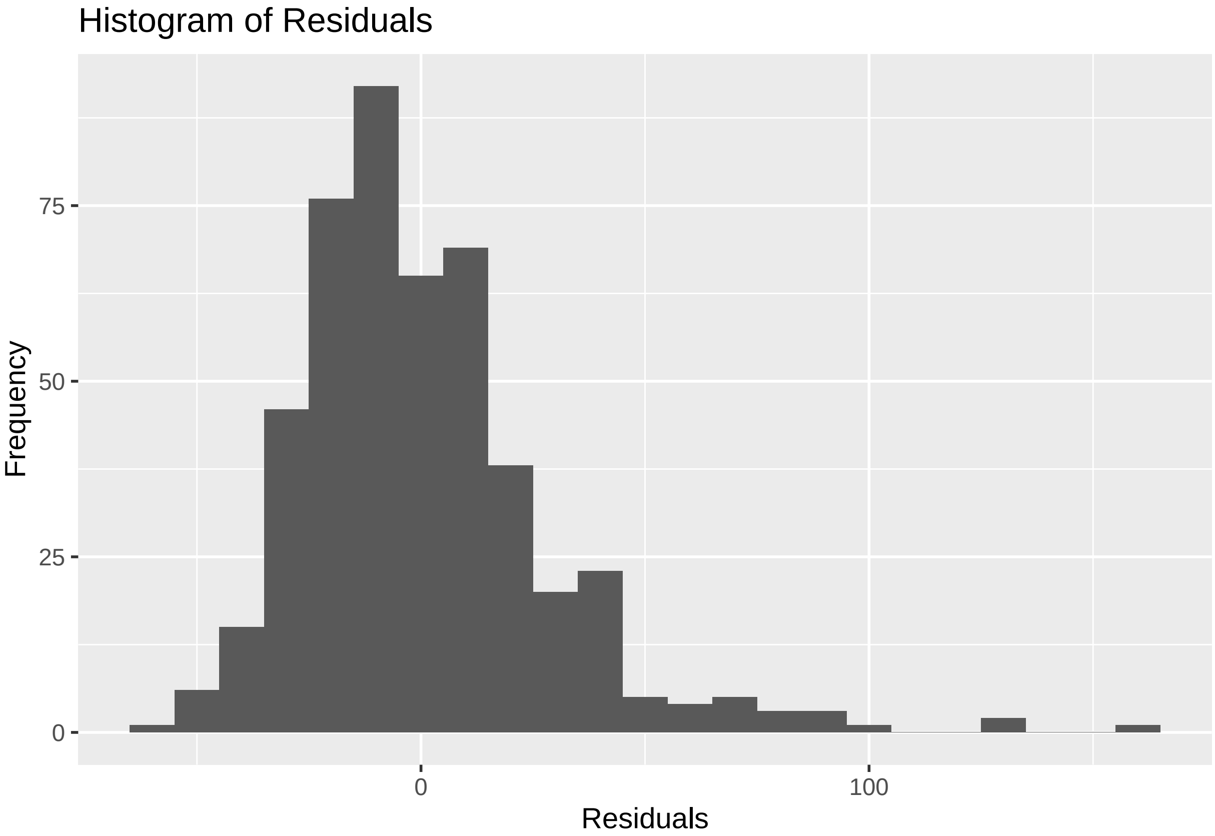
36 Histogram of model residuals

Raw data: residuals plot code

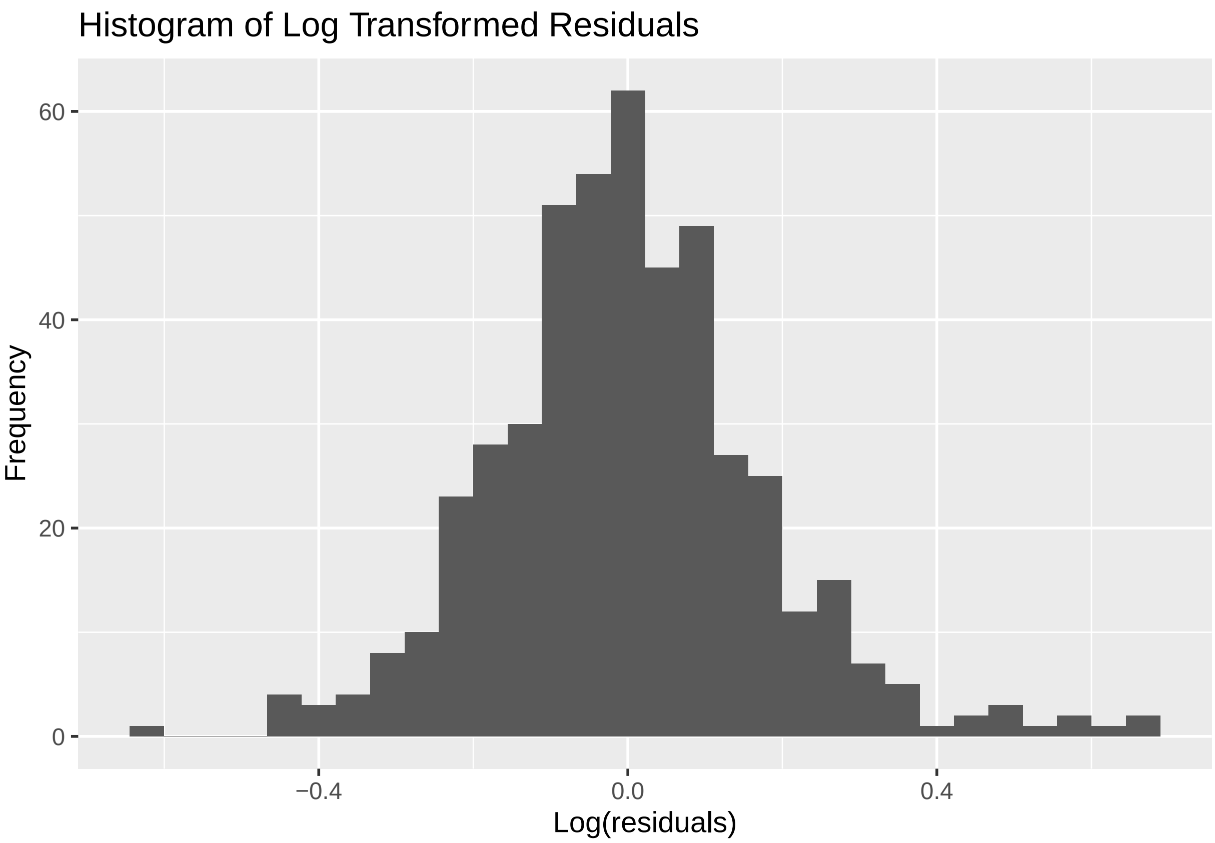
```
# Create model for raw data
mod1 <-
  lm(weight_lbs ~ height_in,
     data = handw)

# Create residuals object
residuals_df <-
  mod1$residuals |>
  as_tibble() |>
  rename(residuals = value)

# Plot the data
ggplot(residuals_df, aes(x = residuals)) +
  geom_histogram(binwidth = 10) +
  labs(title = "Histogram of Residuals",
       x = "Residuals",
       y = "Frequency")
```



37 Raw data: residuals plot output



38 Transformed data: residuals plot

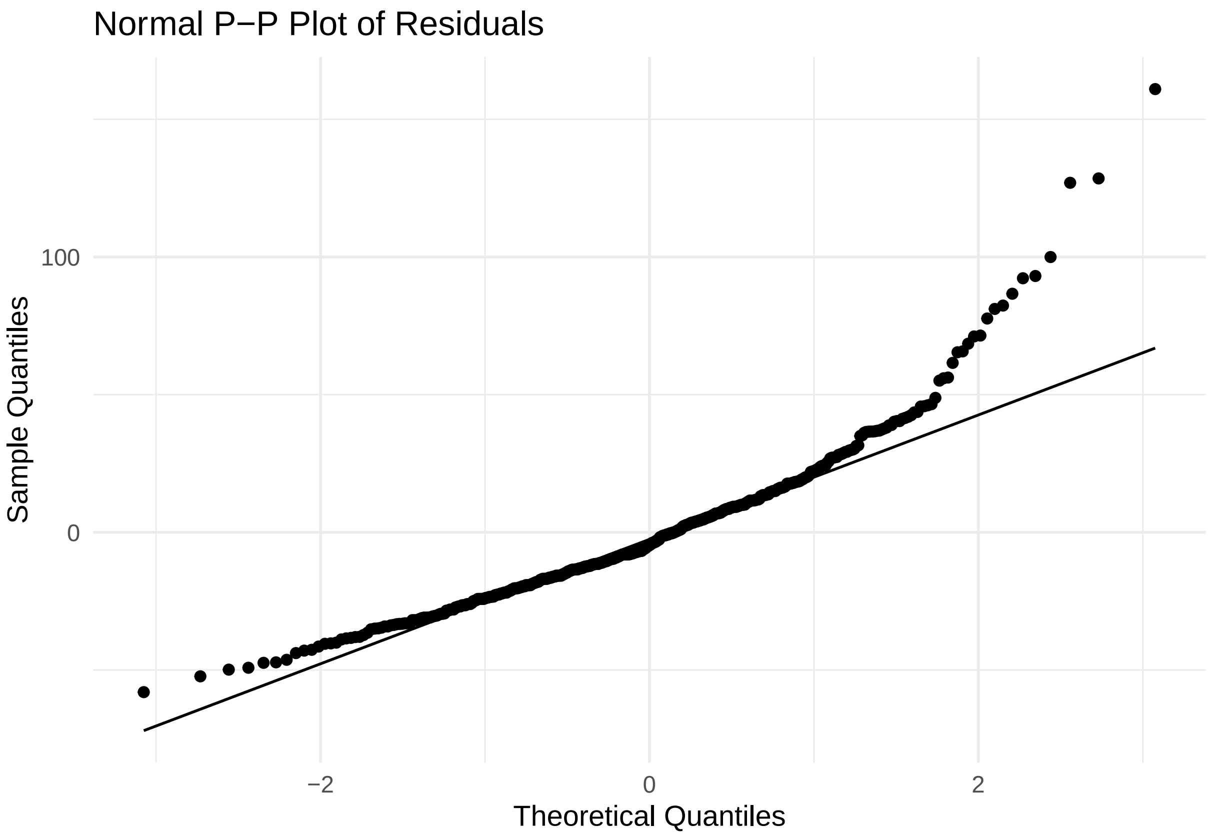


39 Normal Probability-Probability (P-P) Plots

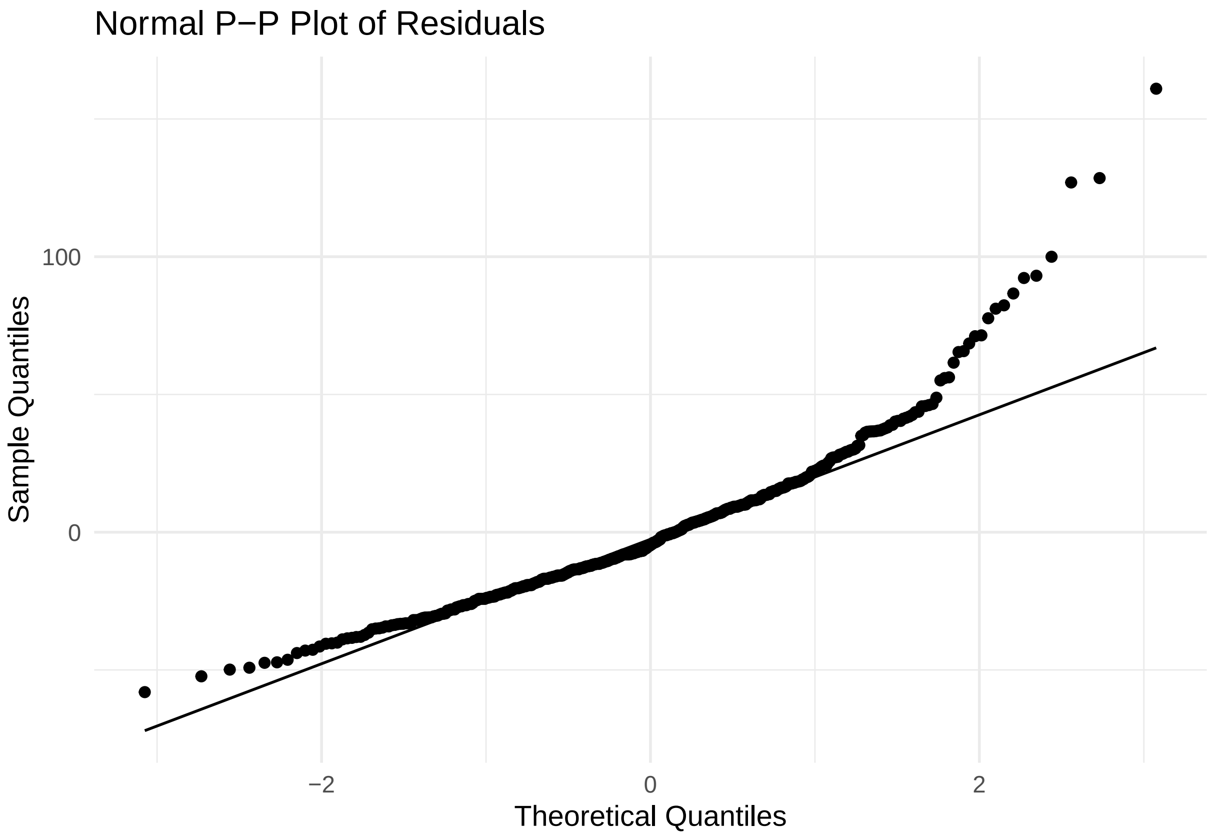
Raw data: Normal PP plot code

```
# Extract residuals
residuals <-
  mod1$residuals

# Calculate theoretical quantiles
mod1_quantiles <-
  qqnorm(residuals, plot.it = FALSE)$x

# Create a data frame with residuals and theoretical quantiles
pp_df <-
  data.frame(
  residuals = residuals,
  theoretical_quantiles = mod1_quantiles
)
```

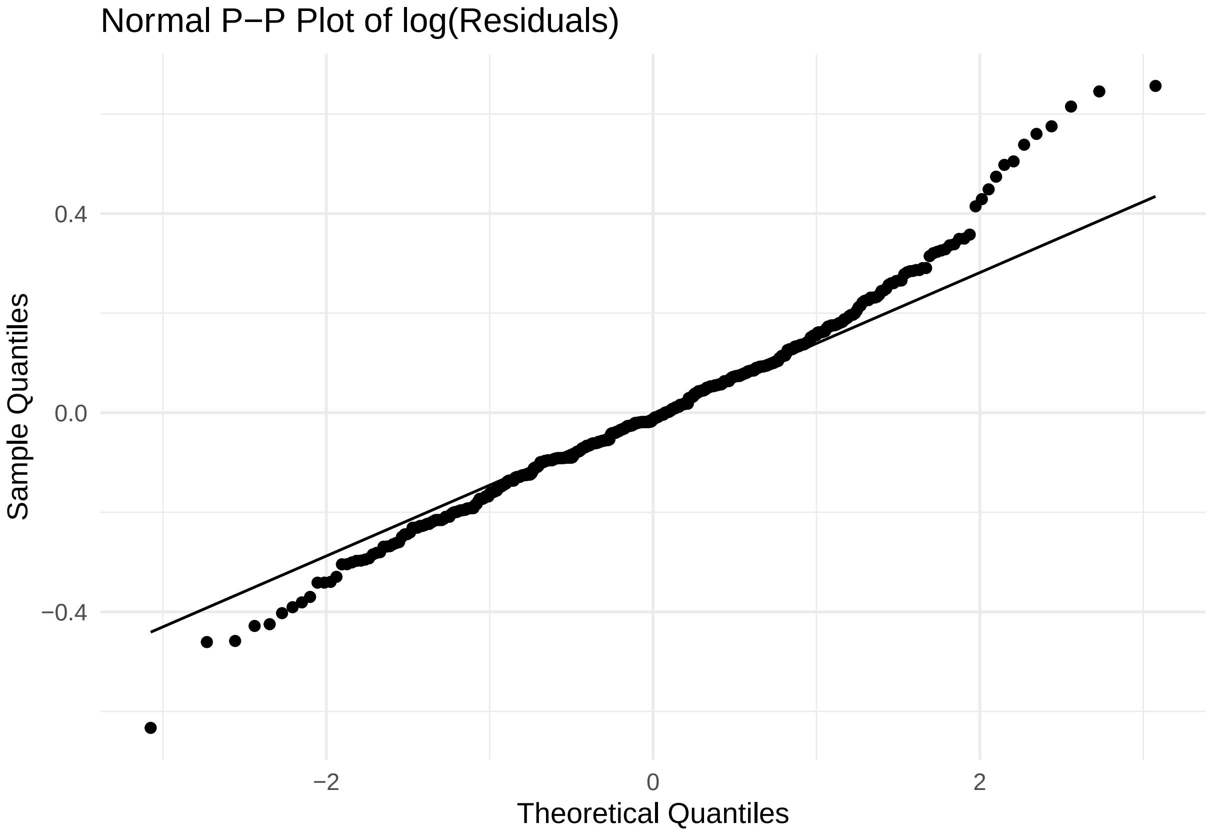
```
# Plot the P-P plot
ggplot(pp_df, aes(sample = residuals)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Normal P-P Plot of Residuals",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles") +
  theme_minimal()
```



40 Raw data: Normal PP plot output



41 Transformed data: Normal PP plot ouput



42 Testing the Significance of the model and its coefficients

We can use statistical tests to determine how well we are approximating the population parameters with those in our model. To do this we can use two methods:

- Analysis of variance (ANOVA): tests overall significance of the model
- t-test: test the individual significance of the coefficients.

42.1 Testing the overall significance of the model

Testing the overall significance of the model evaluates how well the independent variables reliable predict the dependent variable. We can create hypotheses to make our statistical inferences:

H_0: The regression model does not explain a significant proportion of the variance in weight (our DV).

VS

H_1: The regression model does explain a significant proportion of the variation in weight.

And we test the above using the F-distribution.

43 ANOVA for regression modelling

ANOVA for regression

Source of	Sum of Squares	Degrees of Freedom	Mean Square	
Variation	(SS)	(df)	(MS)	F-Statistic (F)
Model	$\text{SS}_{\text{model}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	k	$\text{MSR} = \frac{\text{SSR}}{k}$	$\frac{\text{MSR}}{\text{MSE}}$
Residual	$\text{SS}_{\text{residual}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$n - k - 1$	$\text{MSE} = \frac{\text{SSE}}{n-k-1}$	
Total	$\text{TSS} = SSR + SSE = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$n - 1$		

ANOVA for regression: key

- n = total number of observations
- k = total number of independent variables
- y = the observed values
- $\bar{y}$ = mean of the dependent variables
- $\hat{y}_i$ = the estimated values

44 ANOVA for log transformed model

```
# Run an ANOVA on our log transformed model
anova_results <-
  anova(mod2)

print(anova_results)
```

```
## Analysis of Variance Table
##
## Response: wlog
##             Df  Sum Sq Mean Sq F value                Pr(>F)
## hlog         1 182.622 182.622  5801.8 < 0.00000000000000022 ***
## Residuals 473  14.888   0.031
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Calculate TSS
ss_model <-
  anova_results["hlog", "Sum Sq"]

ss_residual <-
  anova_results["Residuals", "Sum Sq"]

tss <-
  ss_model + ss_residual

print(tss)
```

```
## [1] 197.5102
```

44.1 Interpreting the results of ANOVA

We can see from the output that `hlog` that the slope of the line is significant difference from 0, where $p < 0.001$. Thus we reject the null hypothesis (H_0).

References

Barr, Dale J. 2025. "Learning Statistical Models Through Simulation in r: An Interactive Textbook." In *PsyTeachR*, 1.0.0 ed. Creative Commons. https://psyteachr.github.io/stat-models-v1/.

Canduela, Jesus, and Robert Raeside. 2020. *The Quantitative Researcher*. Heriot-Watt University. www.hw.ac.uk/ebs.