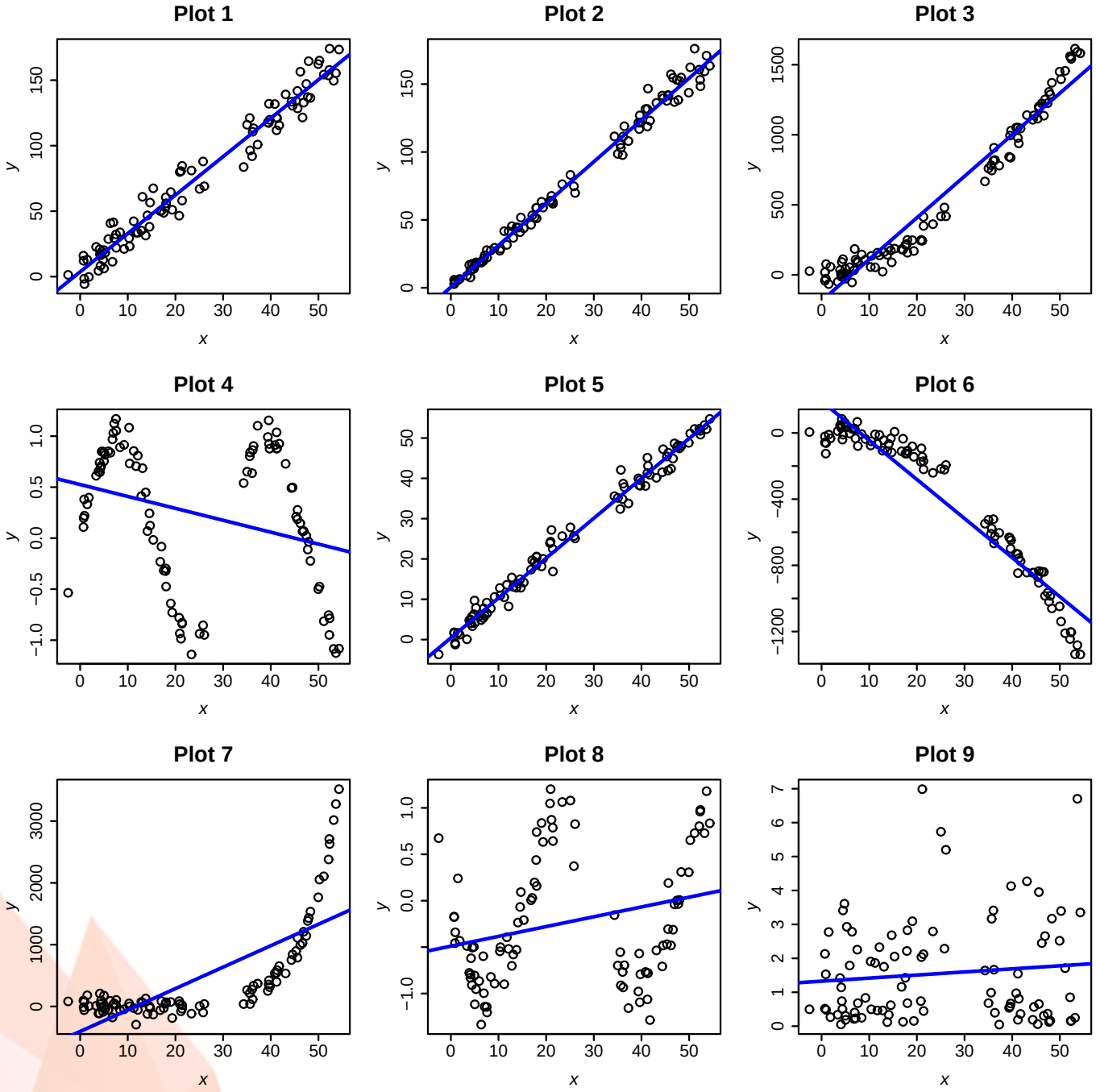
STM4PSD – Workshop 11

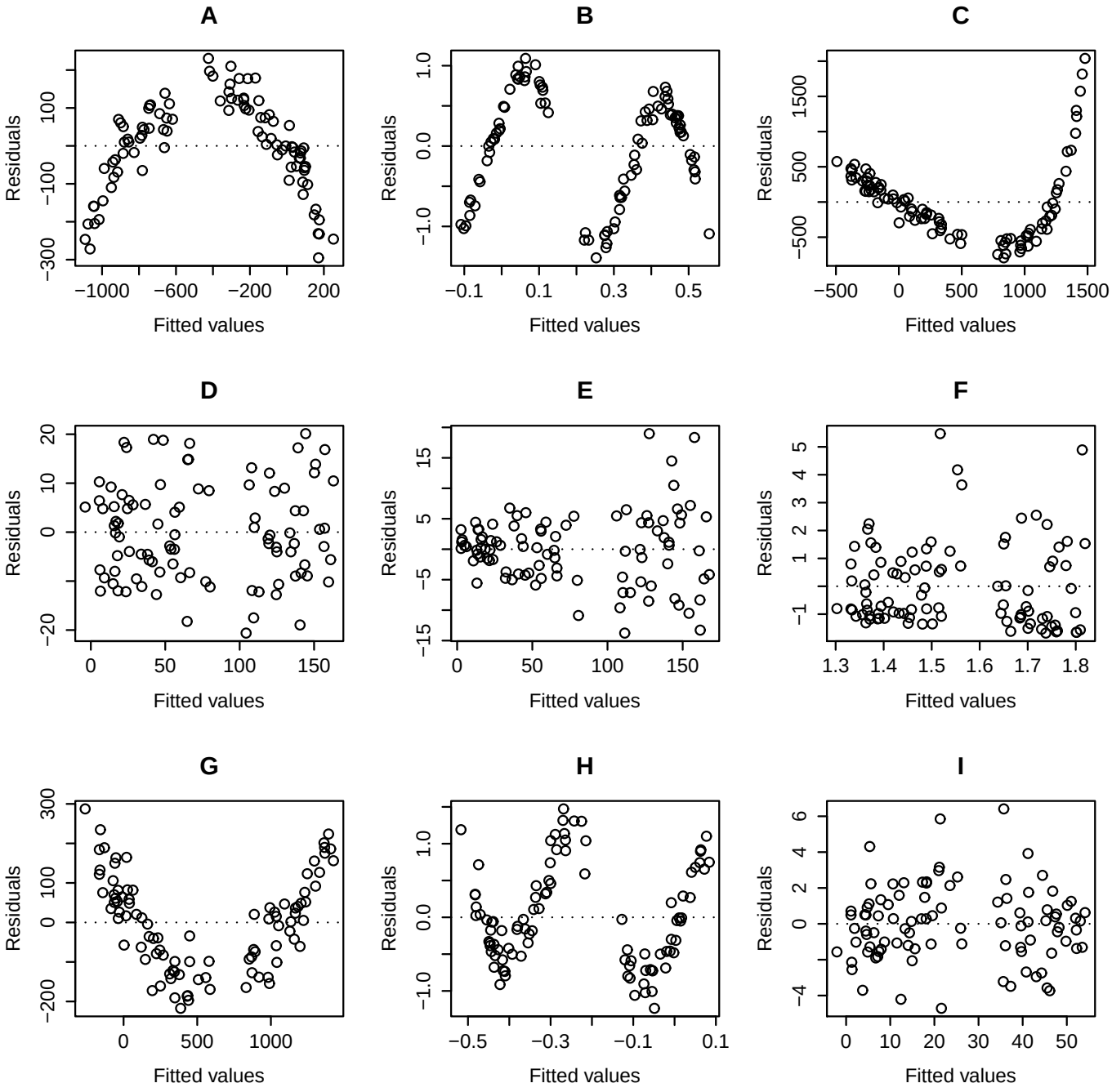
Residuals versus fits plots

1. Below, there are 9 different plots shown, each depicting a data set of size $n = 100$. The x_i values are on the horizontal axis and the y_i values are on the vertical axis, and their estimated least squares line is shown in blue.



(a) Using your knowledge of fits and residuals, try and imagine what the associated residuals versus fits plots would look like for each of these 9 plots.

(b) For which of these plots do you think the simple linear regression model (including the associated error assumption of independent errors with constant variance σ^2) holds at least approximately? Which plots suggest clear violations of the simple linear regression model and which are you not sure about?



(c) The residual versus fits plots for each of the 9 data sets are shown above. However, the order of these plots is random so that, for example, residuals versus fits plot 'A' is not necessarily the plot associated with 'Plot 1' above. Your task is to match the plots in part (a) with their respective residuals versus fits plot. To do so, complete the below table by writing the label under the matching label. One has been done for you already.

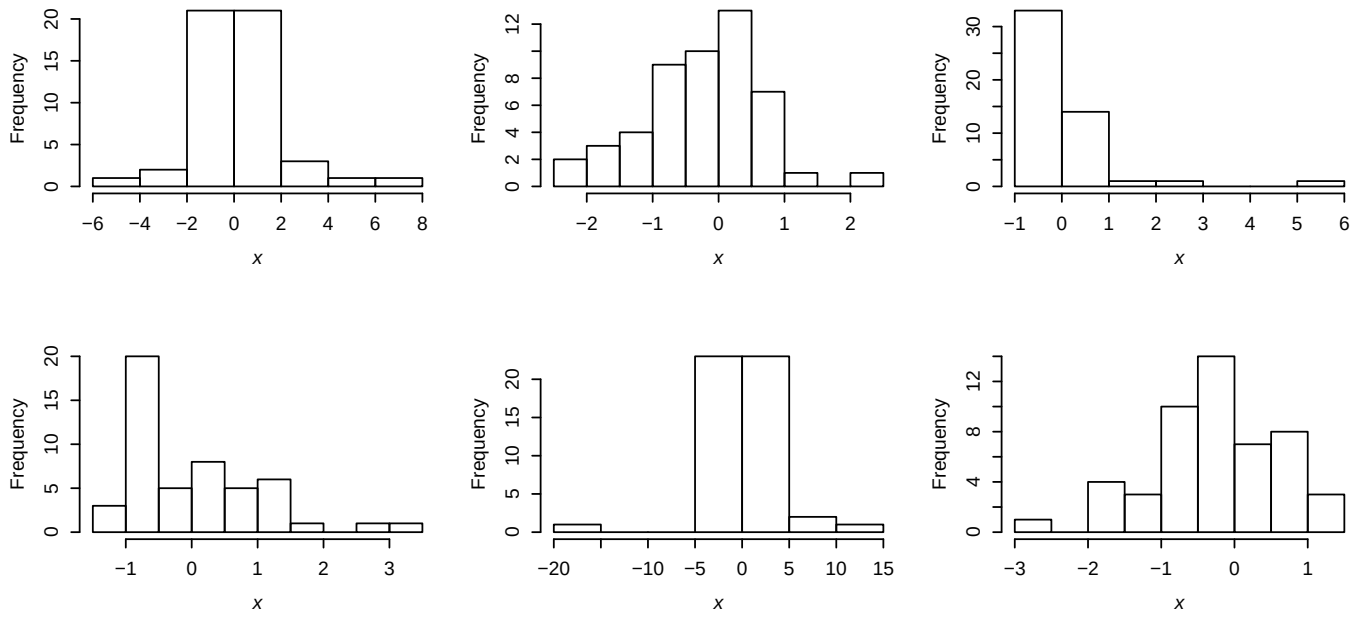
Plot 1	Plot 2	Plot 3	Plot 4	Plot 5	Plot 6	Plot 7	Plot 8	Plot 9
				Plot I				

(d) Using the residual versus fits plots your completed table, are you still happy with your choice of data sets that satisfy the simple linear regression model? What about for those that do not satisfy the simple linear regression model? If you were previously unsure whether some data sets did or didn't satisfy the simple linear regression model, did the residuals versus plots help?

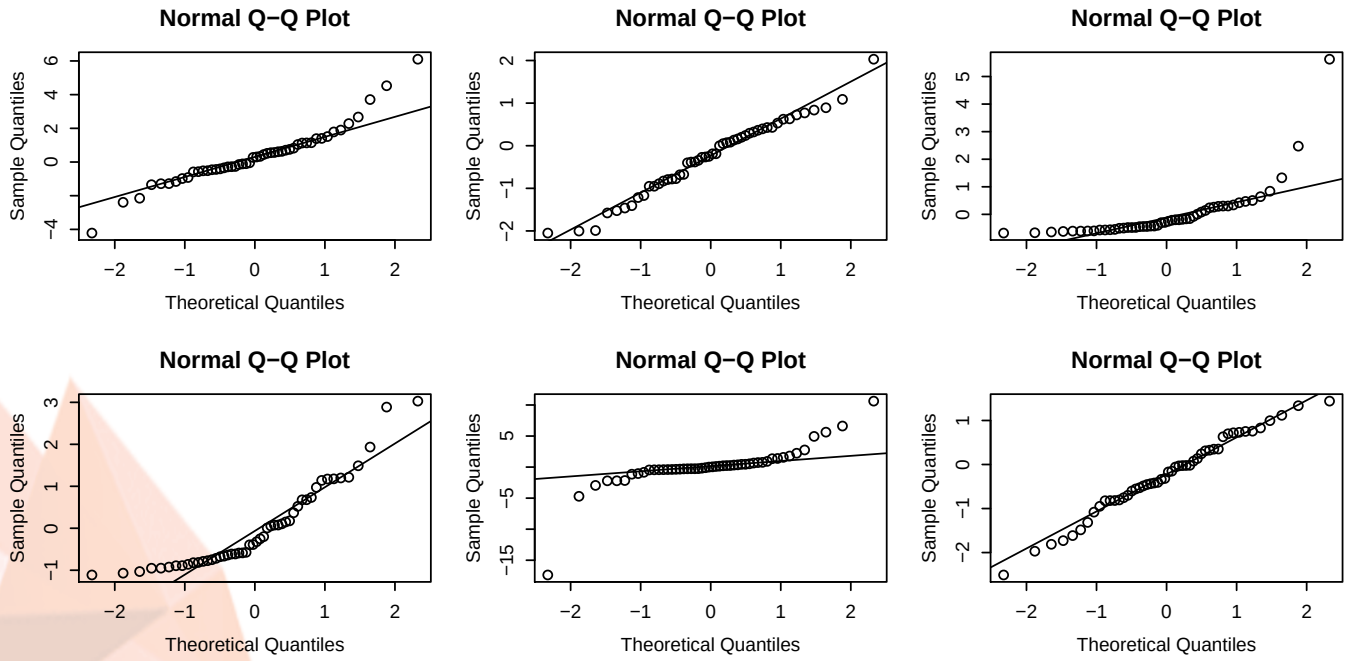
Q-Q plots

Recall that Q-Q plots can be used to check for normality. This is not necessarily only used for regression analysis, but for checking for normality in general. We will start be looking at histograms to check for normality.

2. Below are some histograms for $n = 50$ observations sampled from six distributions. Which ones do you think correspond to data sampled from a normal distribution and why? **Hint**: there are two).



3. For the same data considered in the previous question, the corresponding Q-Q plots are below.



(a) Now use these Q-Q plots to decide on which data corresponds to sampled normal data. State why you chose these.

(b) Based on your experience with these past two questions, which approach do you think is easiest to use to check for violations of normality?

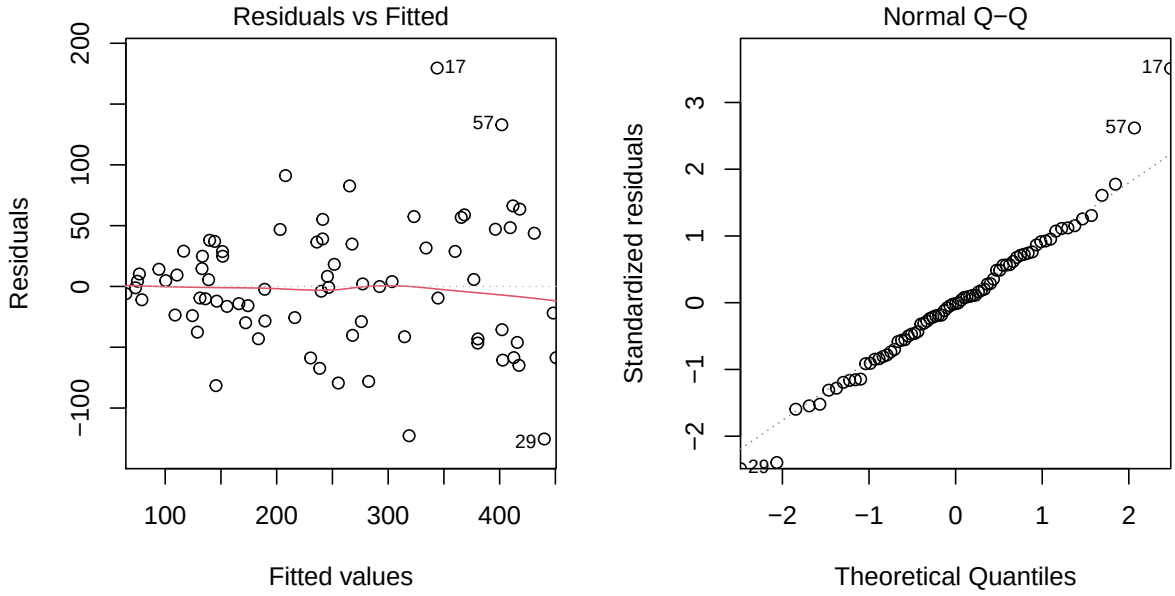
4. In this question, you will be shown some output from a linear regression analysis carried out in R.

Suppose you were applying linear regression to model the amount of toxic substance in a poisonous mushroom, on the basis of its physical attributes. For this question, the explanatory variable is the diameter of the cap of the mushroom (denoted by `diameter`), which is measured in millimetres (mm). The response variable is the amount of toxic substance in the mushroom, measured in micrograms (µg).

Simple linear regression has been carried out in R using the data from 77 poisonous mushrooms. The output is shown below, along with the residuals versus fits plot and a Q-Q plot of the residuals.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.68511   14.03993   0.262    0.794
diameter    4.9864     0.2574   19.372   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 51.73 on 75 degrees of freedom
Multiple R-squared:  0.8334,    Adjusted R-squared:  0.8312
F-statistic: 375.3 on 1 and 75 DF,  p-value: < 2.2e-16
```



(a) Do you think the Residuals versus Fits plot or the Q-Q plot of the residuals suggest that there are any linear regression model violations to be concerned with? Justify your answer clearly with references to both plots.

NOTE: Regardless of your answer to (a), for the remainder of this question, assume that there are no linear regression model violations.

(b) Does the R output suggest that the regression model fits the data well? Explain.

(c) What is the estimate of the explanatory variable coefficient? Interpret this estimated coefficient.

(d) Let β_1 denote the true coefficient for the explanatory variable and consider the hypotheses

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

Do you reject the null hypothesis at the $\alpha = 0.05$ significance level? Explain.

(e) Using the fact that $t_{75,0.975} = 1.992$, construct a 95% confidence interval for β_1.

(f) In the context of the problem, interpret the confidence interval you calculated in part (e).

(g) Suppose we were to predict the amount of toxic substance in a mushroom with a 50 mm cap diameter. The R output for a 95% prediction interval of a mushroom with this cap diameter is shown below.

```
      fit      lwr      upr
1 502.327 395.4237 609.2304
```

Provide a simple, in-context statement that interprets this 95% prediction interval.