Lab 7

Checking Assumptions

We'll now look at a number of tools for checking the assumptions of a linear model. To test these tools, we'll use data simulated from three models:

$$\text{Model 1:} \quad Y = 3 + 5X + \epsilon, \quad \epsilon \sim N(0, 1)$$

$$\text{Model 2:} \quad Y = 3 + 5X + \epsilon, \quad \epsilon \sim N(0, X^2)$$

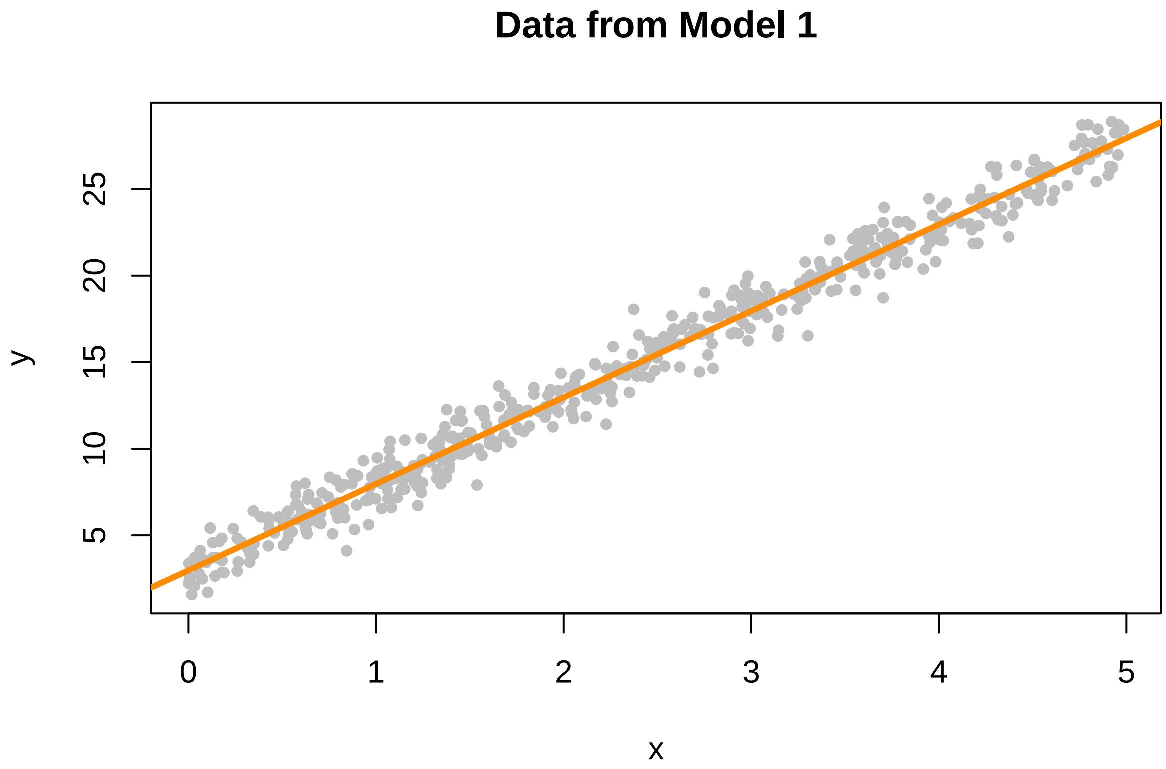
$$\text{Model 3:} \quad Y = 3 + 5X^2 + \epsilon, \quad \epsilon \sim N(0, 25)$$

```r
sim_1 = function(sample_size = 500) {
  x = runif(n = sample_size)*5
  y = 3 + 5 * x + rnorm(n = sample_size, mean = 0, sd = 1)
  data.frame(x, y)
}

sim_2 = function(sample_size = 500) {
  x = runif(n = sample_size)*5
  y = 3 + 5 * x + rnorm(n = sample_size, mean = 0, sd = x)
  data.frame(x, y)
}

sim_3 = function(sample_size = 500) {
  x = runif(n = sample_size)*5
  y = 3 + 5* x ^ 2 + rnorm(n = sample_size, mean = 0, sd = 5)
  data.frame(x, y)
}
```

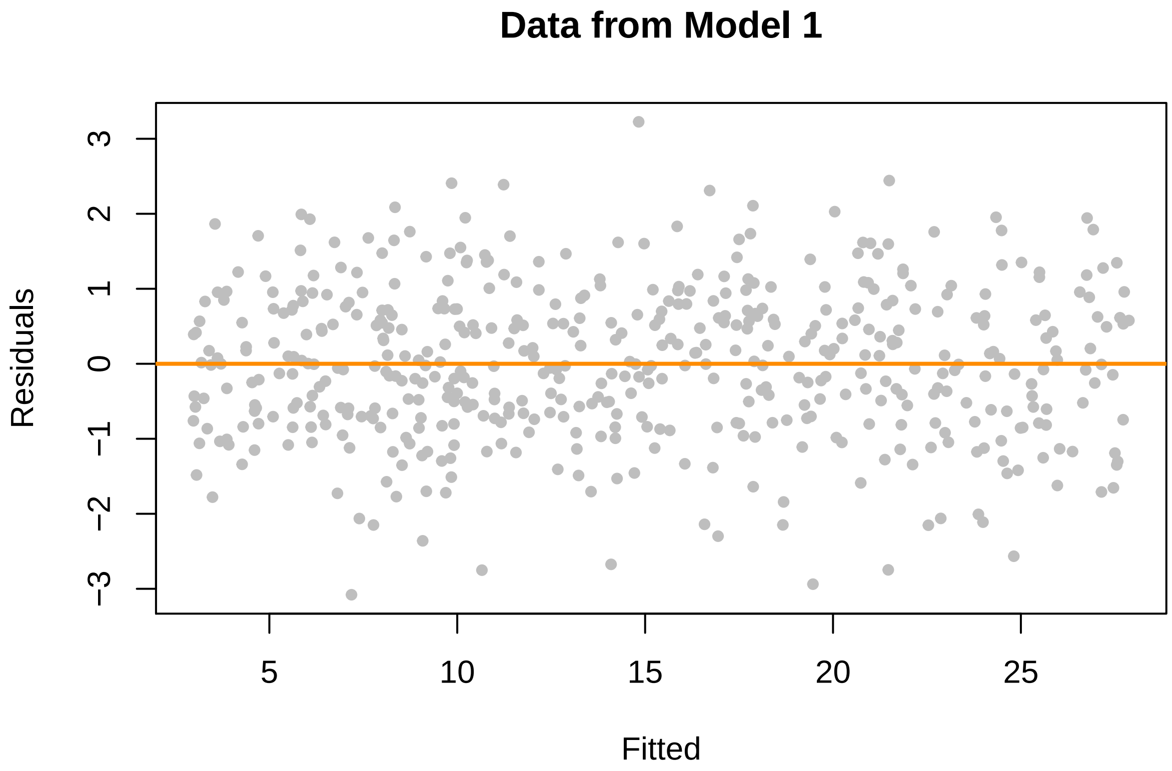
Task 1 (Fitted versus Residuals Plot)

Fit the models and add the fitted lines to scatterplots. Then create for each model case the corresponding residual plots against the fitted values. What do you observe regarding both the linearity and constant variance assumptions?

```r
sim_data_1 = sim_1()
plot(y ~ x, data = sim_data_1, col = "grey", pch = 20,
     main = "Data from Model 1")
fit_1 = lm(y ~ x, data = sim_data_1)
abline(fit_1, col = "darkorange", lwd = 3)
```

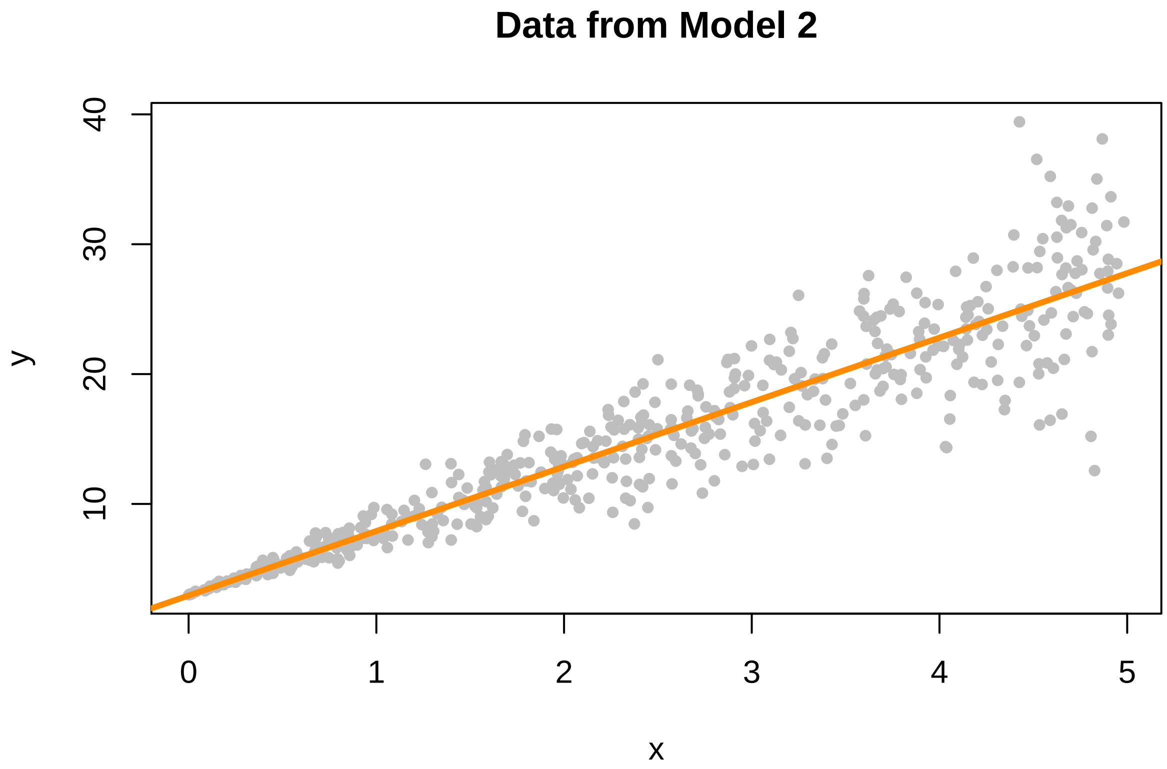


```r
plot(fitted(fit_1), resid(fit_1), col = "grey", pch = 20,
     xlab = "Fitted", ylab = "Residuals", main = "Data from Model 1")
abline(h = 0, col = "darkorange", lwd = 2)
```

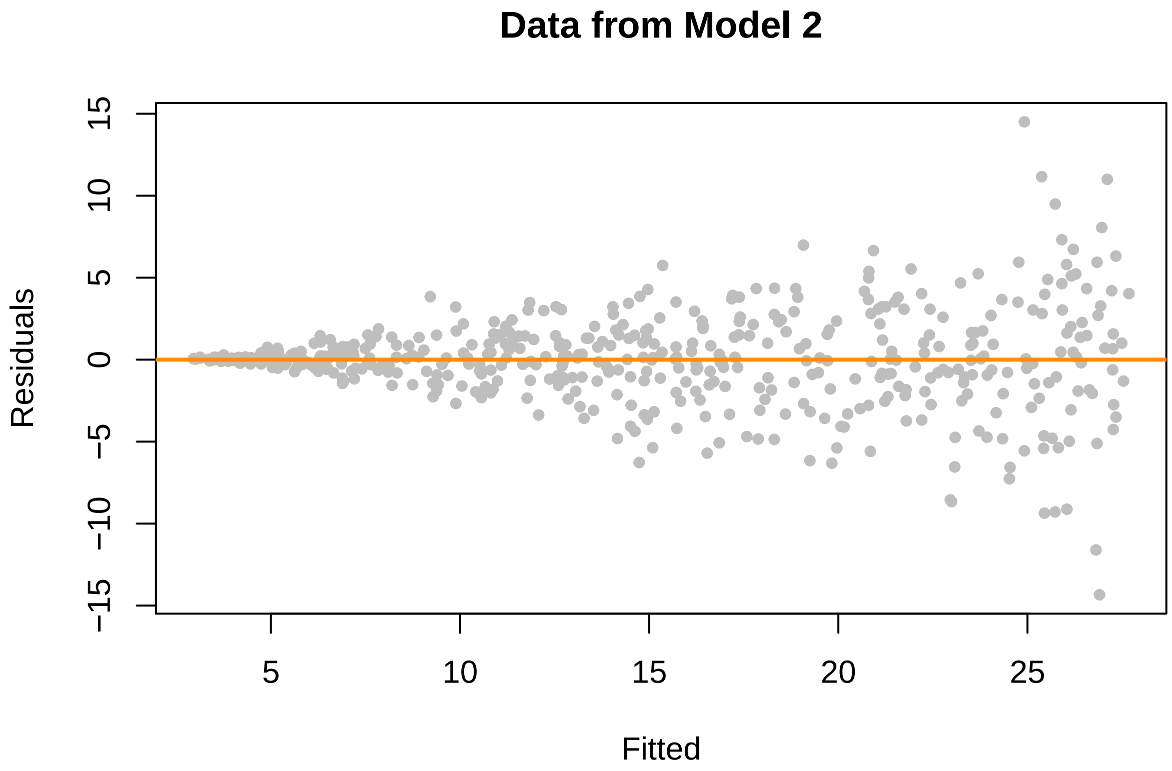


Model 2 is an example of non-constant variance. In this case, the variance is larger for larger values of the predictor variable X.

```
set.seed(42)
sim_data_2 = sim_2()
fit_2 = lm(y ~ x, data = sim_data_2)
plot(y ~ x, data = sim_data_2, col = "grey", pch = 20,
     main = "Data from Model 2")
abline(fit_2, col = "darkorange", lwd = 3)
```



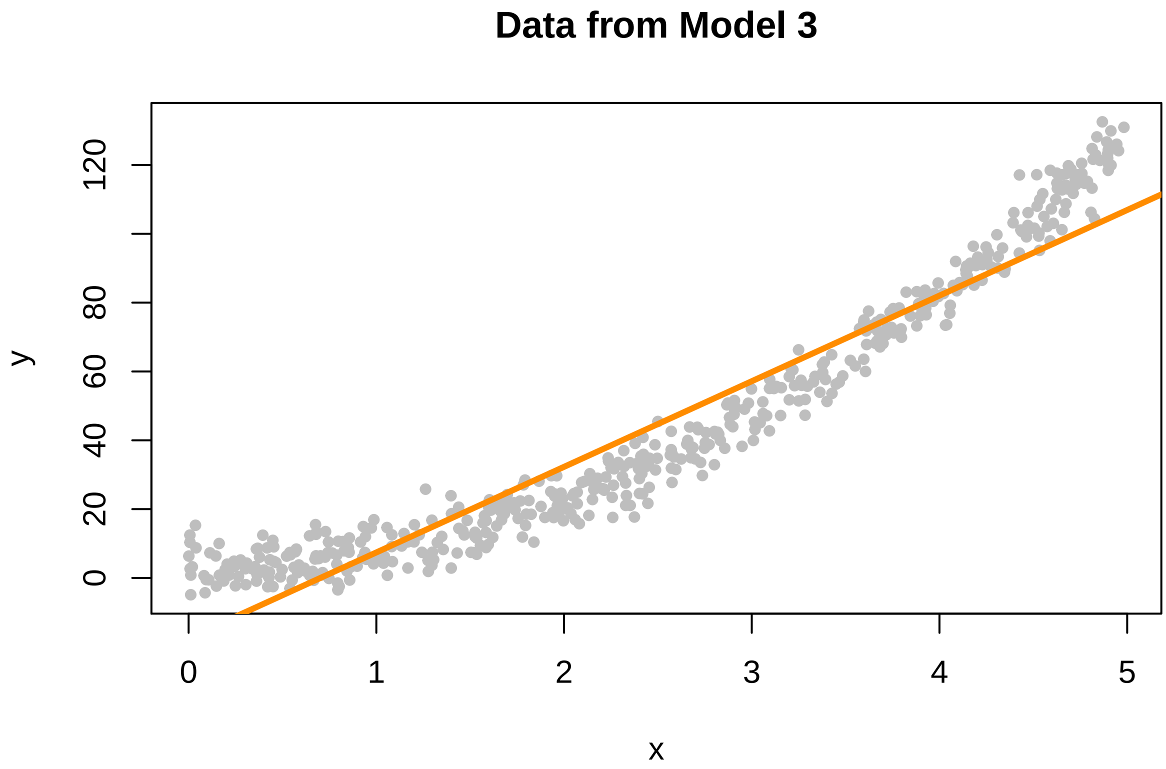
```r
plot(fitted(fit_2), resid(fit_2), col = "grey", pch = 20,
     xlab = "Fitted", ylab = "Residuals", main = "Data from Model 2")
abline(h = 0, col = "darkorange", lwd = 2)
```



On the fitted versus residuals plot, we see two things very clearly. For any fitted value, the residuals seem roughly centered at 0. This is good! The linearity assumption is not violated. However, we also see very clearly, that for larger fitted values, the spread of the residuals is larger. This is bad! The constant variance assumption is violated here.

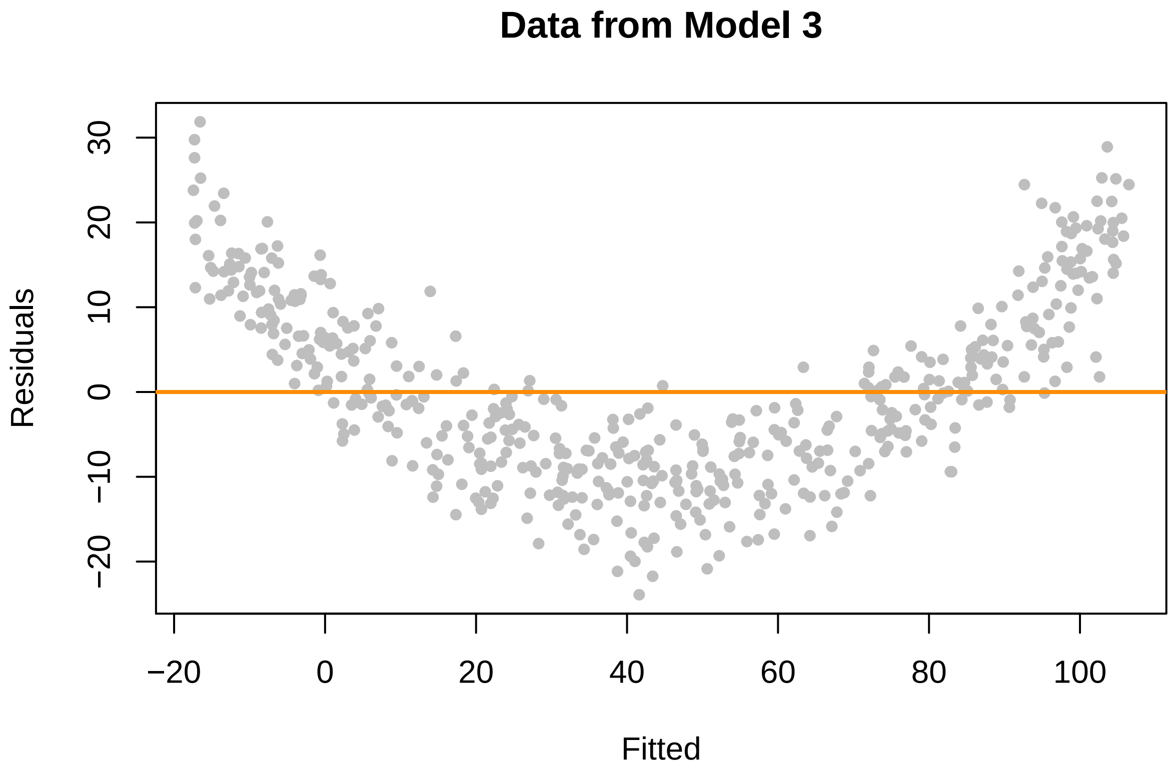
Now we will demonstrate a model which does not meet the linearity assumption. Model 3 is an example of a model where Y is not a linear combination of the predictors. In this case the predictor is X, but the model uses X^2. (We'll see later that this is something that a "linear" model can deal with. The fix is simple, just make X^2 a predictor!)

```r
set.seed(42)
sim_data_3 = sim_3()
fit_3 = lm(y ~ x, data = sim_data_3)
plot(y ~ x, data = sim_data_3, col = "grey", pch = 20,
     main = "Data from Model 3")
abline(fit_3, col = "darkorange", lwd = 3)
```



Again, this is rather clear on the scatterplot, but again, we wouldn't be able to check this plot for multiple regression.

```
plot(fitted(fit_3), resid(fit_3), col = "grey", pch = 20,
     xlab = "Fitted", ylab = "Residuals", main = "Data from Model 3")
abline(h = 0, col = "darkorange", lwd = 2)
```

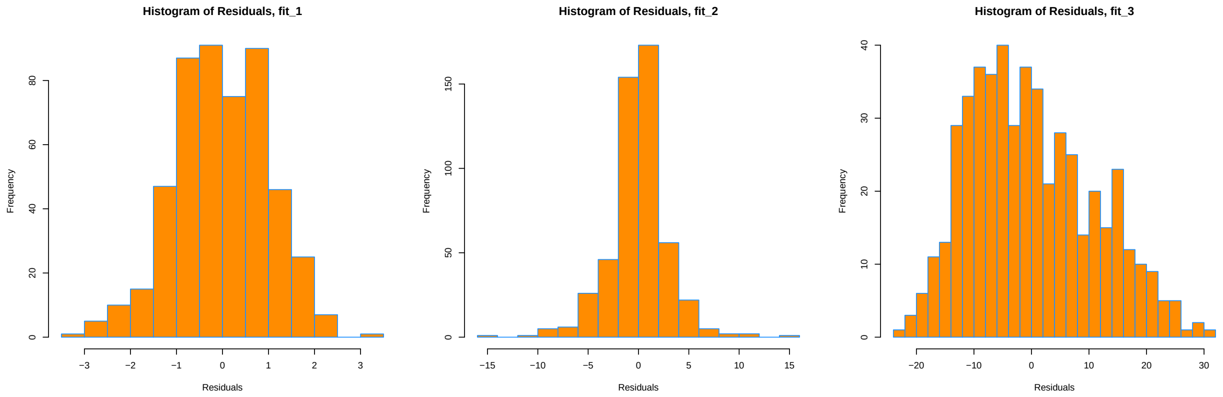


This time on the fitted versus residuals plot, for any fitted value, the spread of the residuals is about the same. However, they are not even close to centered at zero! At small and large fitted values the model is underestimating, while at medium fitted values, the model is overestimating. These are systematic errors, not random noise. So the constant variance assumption is met, but the linearity assumption is violated. The form of our model is simply wrong. We're trying to fit a line to a curve!

Task 2 (Histograms)

For each model case use a histogram of the residuals for assessing the normality assumption. What do you observe regarding normality?

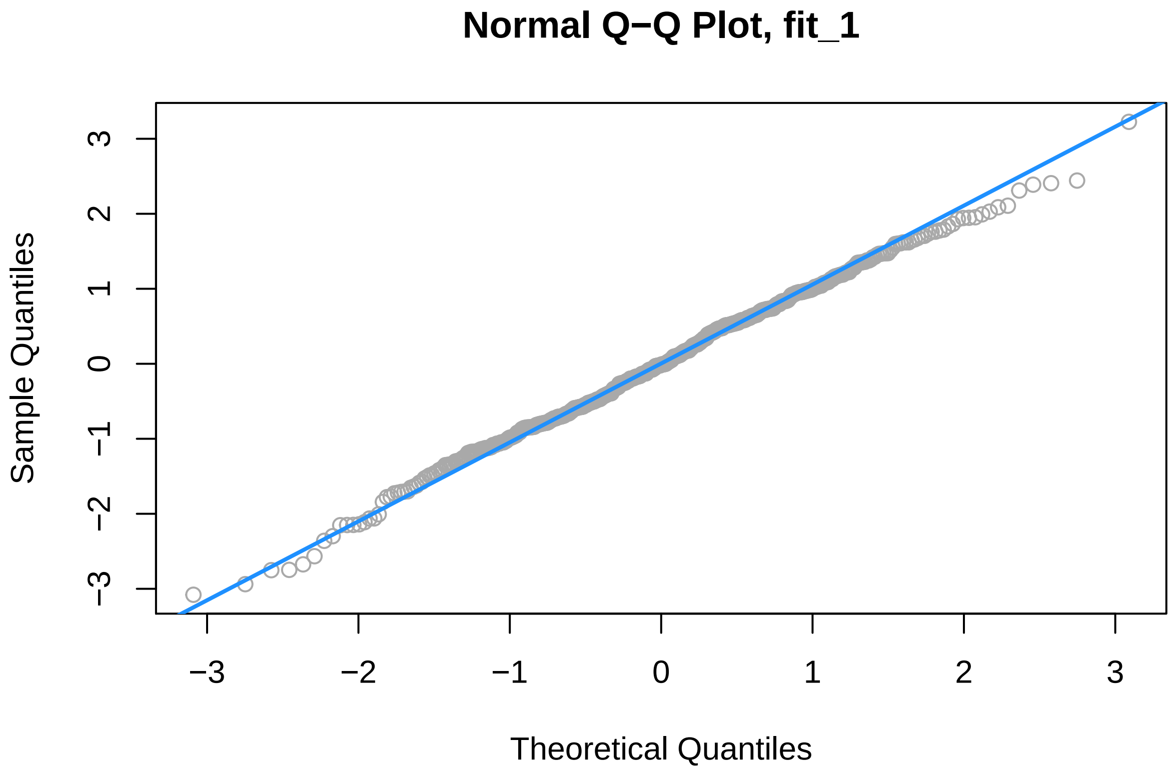
```r
par(mfrow = c(1, 3))
hist(resid(fit_1),
     xlab   = "Residuals",
     main   = "Histogram of Residuals, fit_1",
     col    = "darkorange",
     border = "dodgerblue",
     breaks = 20)
hist(resid(fit_2),
     xlab   = "Residuals",
     main   = "Histogram of Residuals, fit_2",
     col    = "darkorange",
     border = "dodgerblue",
     breaks = 20)
hist(resid(fit_3),
     xlab   = "Residuals",
     main   = "Histogram of Residuals, fit_3",
     col    = "darkorange",
     border = "dodgerblue",
     breaks = 20)
```



Task 3 (Q-Q Plots)

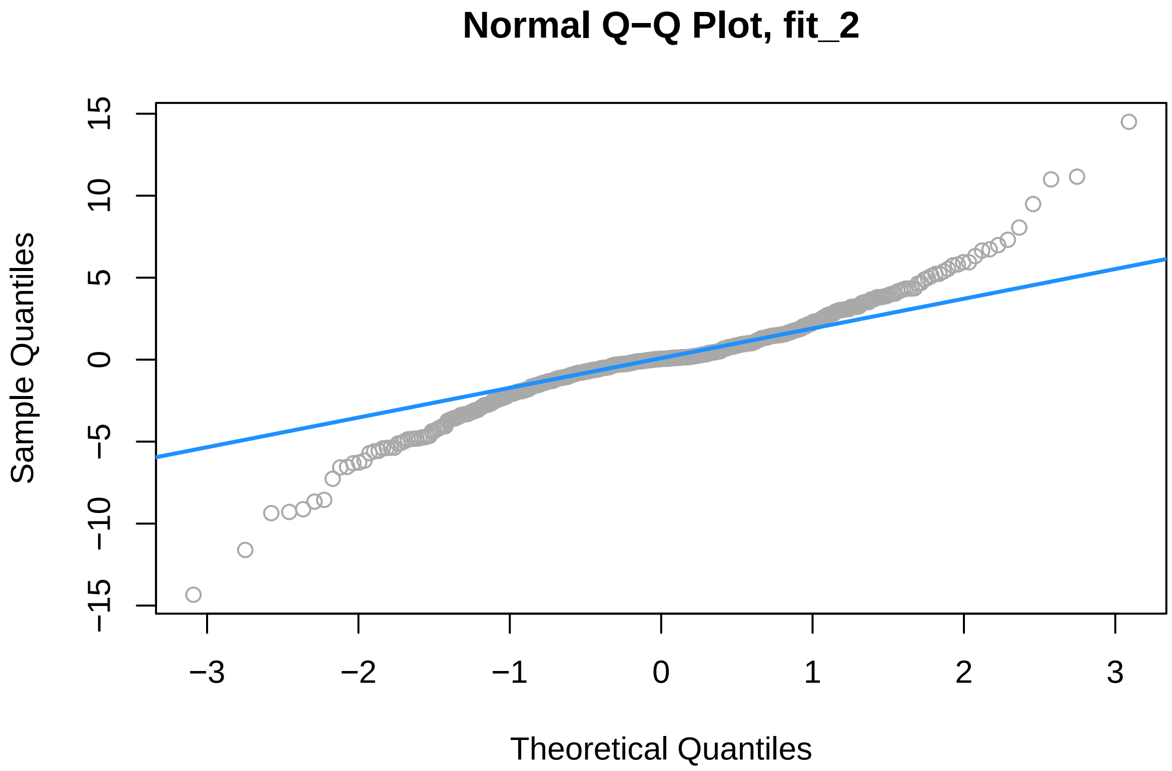
For each model case use a Q-Q plot for assessing the normality assumption. What do you observe regarding normality?

```
qqnorm(resid(fit_1), main = "Normal Q-Q Plot, fit_1", col = "darkgrey")
qqline(resid(fit_1), col = "dodgerblue", lwd = 2)
```



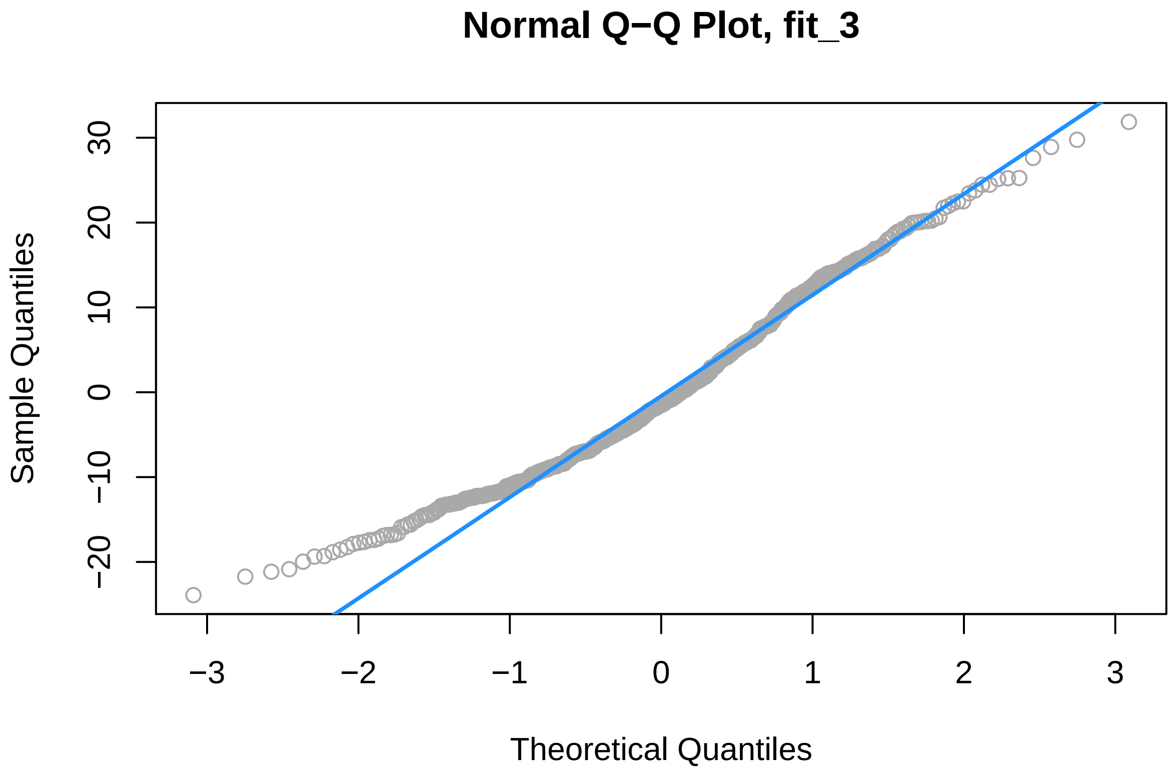
For `fit_1`, we have a near perfect Q-Q plot. We would believe the errors follow a normal distribution.

```
qqnorm(resid(fit_2), main = "Normal Q-Q Plot, fit_2", col = "darkgrey")
qqline(resid(fit_2), col = "dodgerblue", lwd = 2)
```



For `fit_2`, we have a suspect Q-Q plot. We would probably **not** believe the errors follow a normal distribution.

```r
qqnorm(resid(fit_3), main = "Normal Q-Q Plot, fit_3", col = "darkgrey")
qqline(resid(fit_3), col = "dodgerblue", lwd = 2)
```



Lastly, for `fit_3`, we again have a suspect Q-Q plot. We would probably **not** believe the errors follow a normal distribution.