Model Diagnostics: Checking Assumptions!

Often the assumptions of linear regression are stated as:

1. Linearity: The response can be written as a lin. comb. of the predictors.
2. Independence: The errors are indep.
3. Normality: The dist. of the errors should follow a normal dist.
4. Equal Variance: The error variance is the same at any set of predictor values.

Lets look at a number of tools for checking model assumptions by simulating data from three models.

Model 1: $Y = 3 + 5x + \varepsilon, \quad \varepsilon \sim N(0, 1)$

Model 2: $Y = 3 + 5x + \varepsilon, \quad \varepsilon \sim N(0, x^2)$

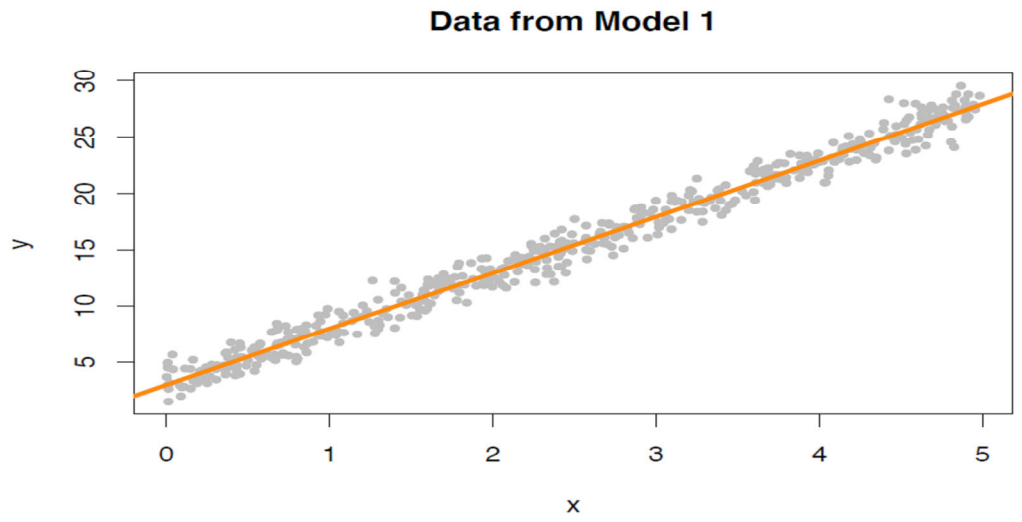
Model 3: $Y = 3 + 5x^2 + \varepsilon, \quad \varepsilon \sim N(0, 25)$

```
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)
}
```

Simulate observations from Model 1 and plot residuals vs. fitted values.
(should ideally fit all assumptions)

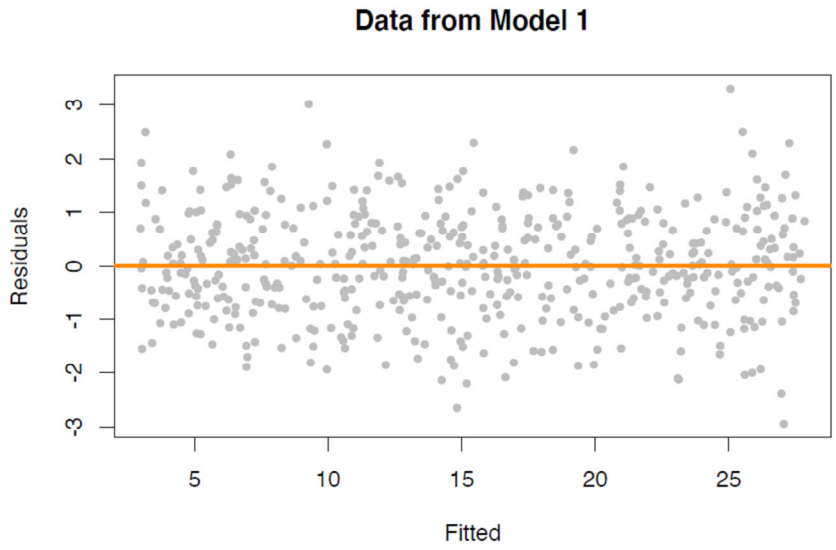
```
set.seed(42)
sim_data_1 = sim_1()
head(sim_data_1)
```

```r
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)
```



Residuals vs. fitted values:

```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)
```



Two things to look for in this plot:

1. At any fitted value, the mean of the residuals

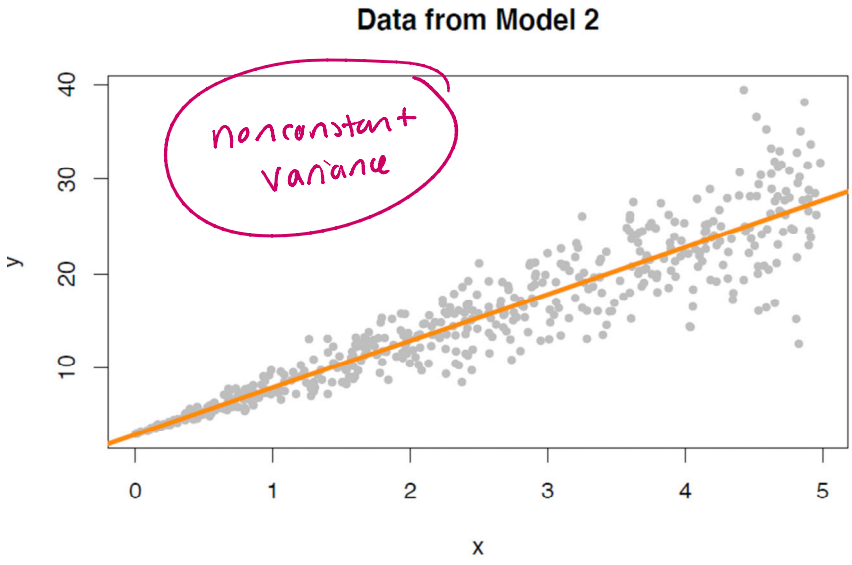
1. At any fitted value, the mean of the residuals should be roughly 0. If this is the case, the linearity assumption is valid.

2. At every fitted value, the spread of the residuals should be roughly the same. If this is the case, the constant variance assumption is valid.

Your turn!!

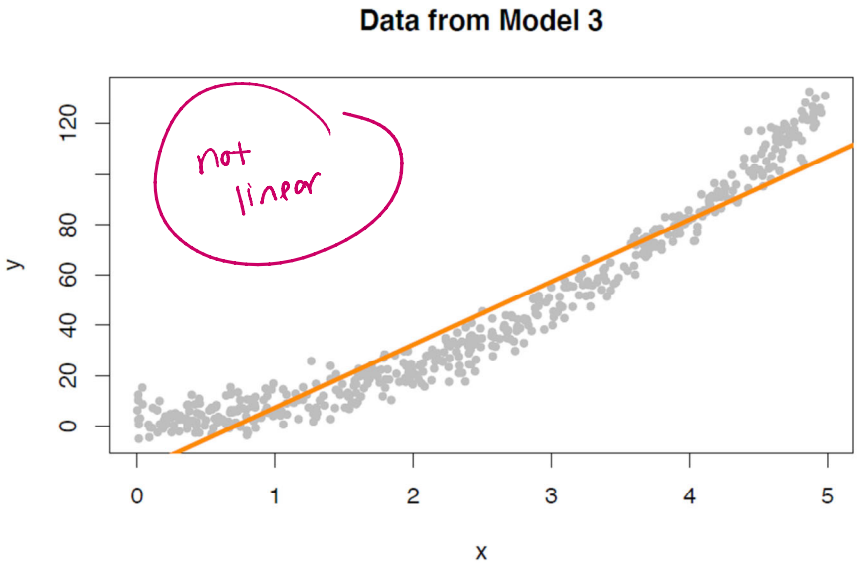
1. Simulate data and generate corresponding plots for Mode 2 and 3.

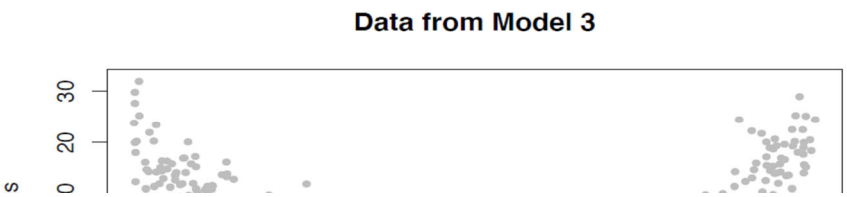
2. What conditions are validated? Not validated?

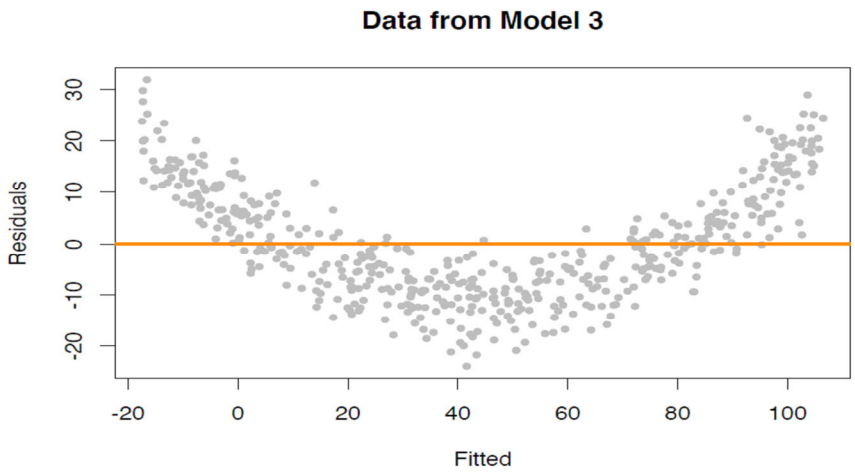



```
plot(fitted(fit_2), resid(fit_2), col = "grey", pch = 20,
     xlab = "Fitted", ylab = "Residuals", main = "Data from Model 2")
abline(h = 0, col = "darkorange", lwd = 2)
```

1. Resids are roughly centered around 0 (yay!)
2. For larger values, the spread of the residuals is larger (boo!)



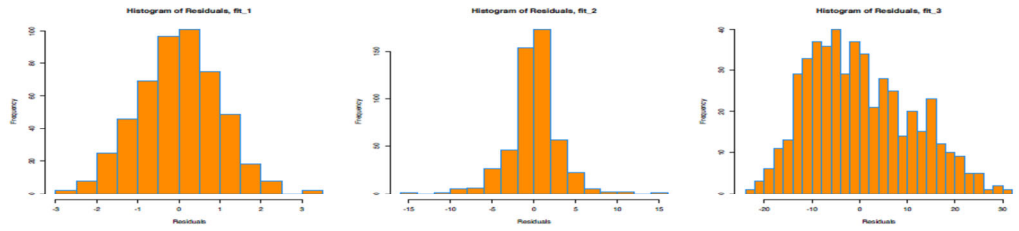




1. Residuals are not centered around 0 (boo!)
2. The spread of the residuals is roughly the same. (yay!)

Histogram: This is a great tool to visually check the normality assumption.

```
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)
```



We will tend to prefer more powerful tools such as QQ Plots and Shapiro-Wilkes test.

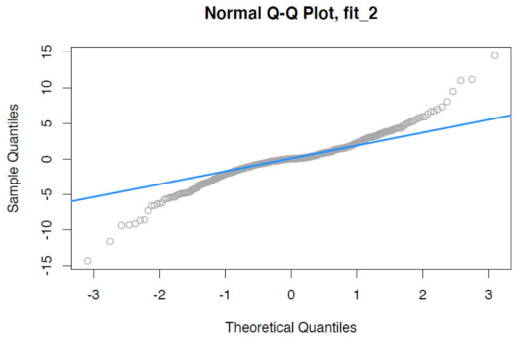
Sorted data

of a normal dist.

- the points don't closely follow a straight line this would suggest they don't come from a normal dist.

Your turn!

1. Generate QQ plots for random normal data of various sizes (10, 25, 100). How do the plots change?
2. Generate QQ plots for random exponential data of various sizes (10, 25, 100). How do these plots change?
2. Create QQ plots for Model 2 and Model 3. Comment on their characteristics.




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

Shapiro - Wilk Test: A formal test of normality.

H_0: The data were sampled from a normal dist.

H_a: The data were not sampled from a normal dist.

Small p-values indicate evidence against normality.

Your turn! Perform the Shapiro-Wilks test on our three models.

```
shapiro.test(resid(fit_1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit_1)
## W = 0.99858, p-value = 0.9622
```

```
shapiro.test(resid(fit_2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit_2)
## W = 0.93697, p-value = 1.056e-13
```

```
shapiro.test(resid(fit_3))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit_3)
## W = 0.97643, p-value = 3.231e-07
```