GENERAL LINEAR MODEL DONE?

LAST SEMESTER QUIZ

Assumptions of the general linear model.

1. Validity.
2. Representativeness.
3. Additivity / linearity.
4. Independence of errors.
5. Equal variance of errors.
6. Normality of errors.

VALIDITY/REPRESENTATIVENESS

Construct validity: is basically a question of whether you're measuring what you want to be measuring. I'm trying to investigate the rates with which university students cheat on their exams.

- I ask the class.
- Everyone says they don't cheat. Can I conclude that no one cheats? **NO!!**
- ***What I really measure:***
  - *"the proportion of people stupid enough to own up to cheating, or bloody minded enough to pretend that they do"*
  - Bad construct validity....

Ecologically validity: the entire set up of the study should closely approximate the real-world scenario that is being investigated.

External validity: relates to the **generalizability** of your findings.

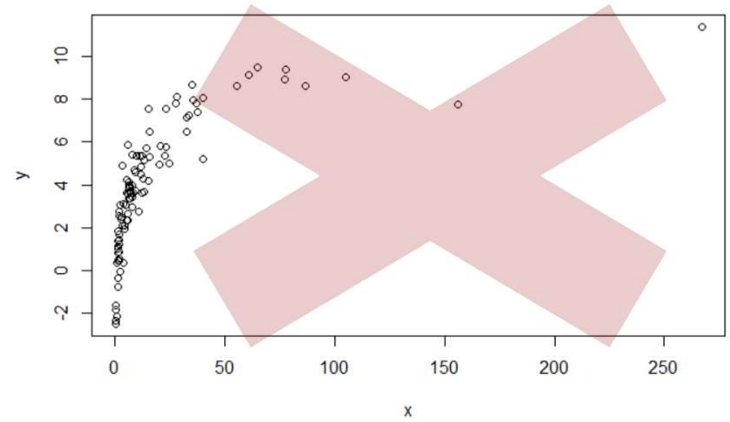
ADDITIVITY / LINEARITY

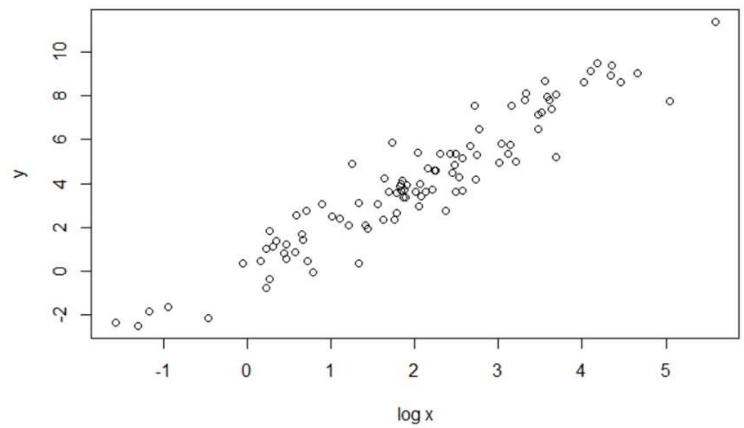
linearity and additivity: of the relationship between dependent and independent variables:

(a) The expected value of dependent variable is a straight-line function of each independent variable, holding the others fixed.

(b) The slope of that line does not depend on the values of the other variables. (Think of adding interactions effects)

(c) The effects of different independent variables on the expected value of the dependent variable are additive.

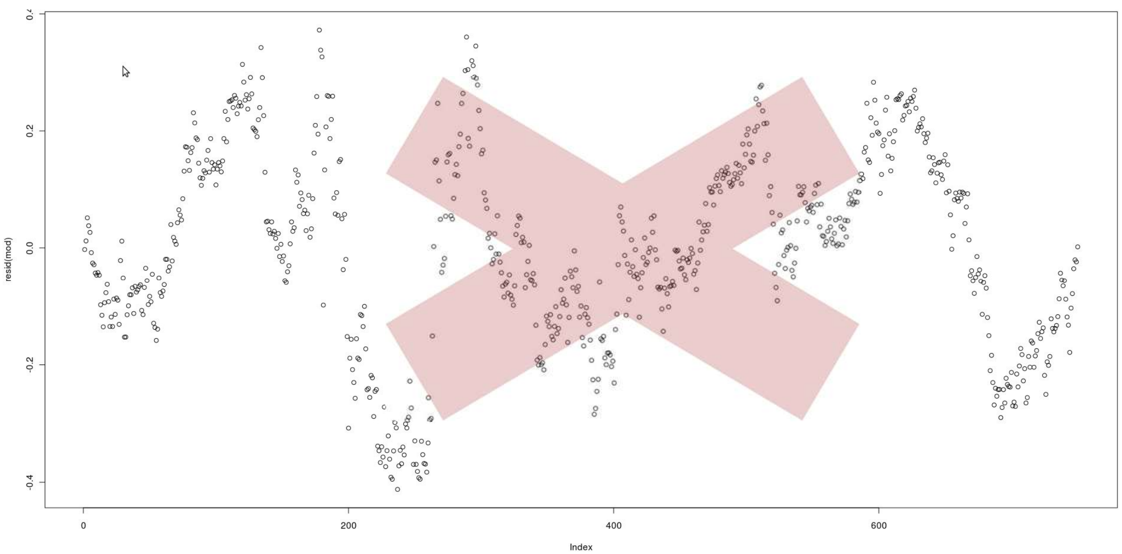




INDEPENDENCE OF ERRORS

The residuals should be randomly and symmetrically distributed around zero under all conditions.

Autocorrelation: is a correlation coefficient. However, instead of correlation between two different variables, the correlation is between two values of the same variable at times X_i and X_{i+k}.



X-axis:

- Time-series (time)
- Row number dependent on in independent variables.

Autocorrelation formula:

$$r_k = \frac{\sum_{t=k+1}^{n} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{n} (y_t - \bar{y})^2}$$

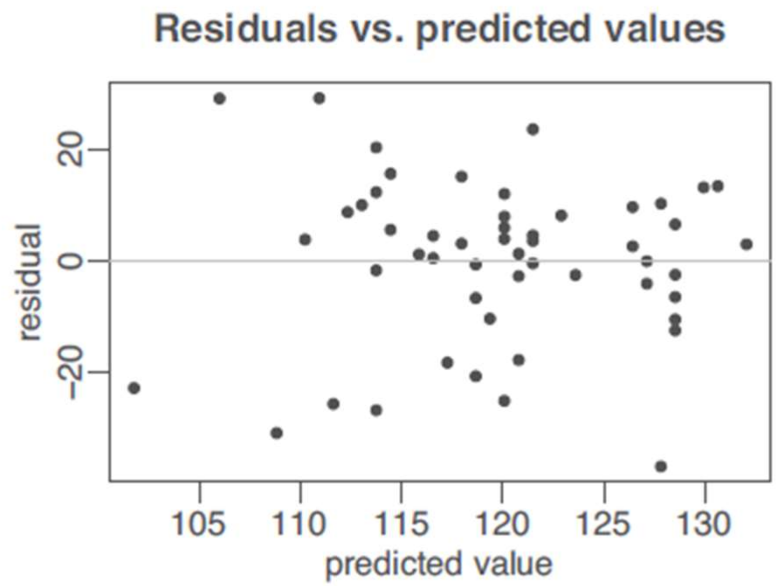
where r_k is the autocorrelation for lag k.

EQUAL VARIANCE OF ERRORS.

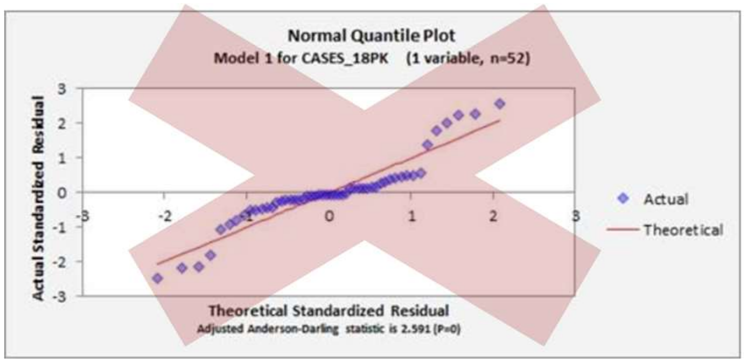
Also called violations of homoscedasticity (Heteroscedasticity).

x = predicted y.

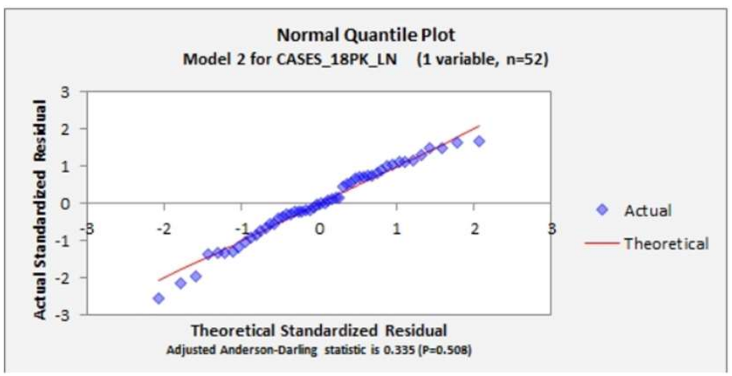
- $\sigma|x_1 = \sigma|x_2 = \sigma|x_n$
- The spread of residuals should be equal for every x.



NORMALITY OF ERRORS



...and here is an example of a good-looking one (a linear pattern with P=0.5 for the A-D stat, indicating no significant departure from normality):



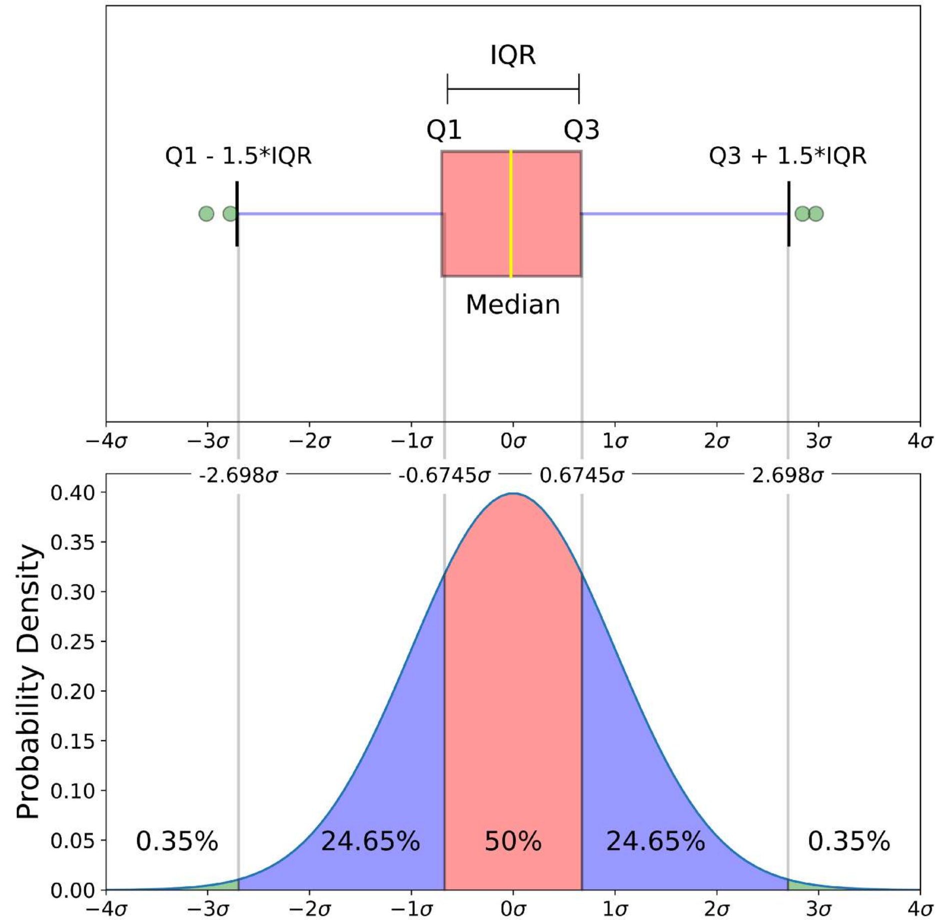
Q-Q PLOTS.

Quantile – Quantile (Q-Q) plots

- What and how?!

A sample is divided into equal-sized, adjacent, subgroups.

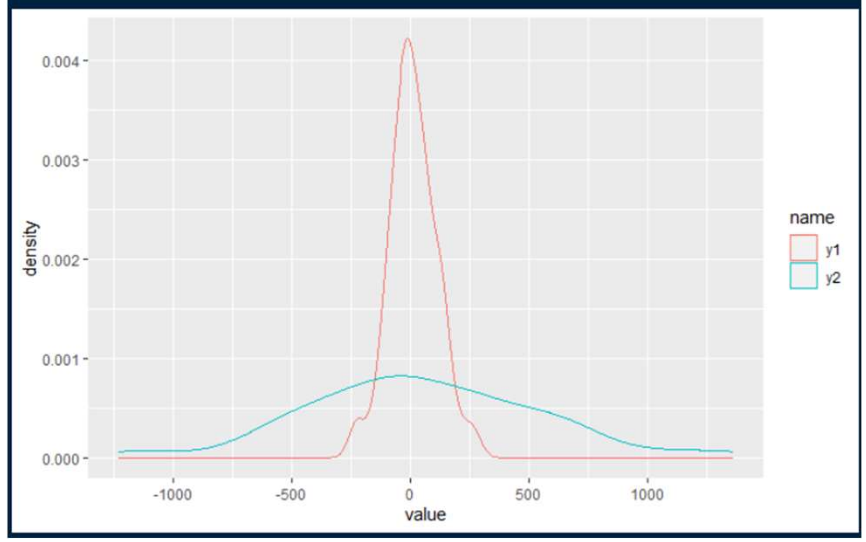
- 2-Quantile: Median (two equal parts)
- 3-Quantile: Tertiles (Three equal parts)
- 4-Quantile: Quartile (Four equal parts)
- And so forth…



Q-Q PLOTS

- Statistics, Q-Q(quantile-quantile) plots play a very vital role to graphically analyze and compare two probability distributions by plotting their quantiles against each other.
  - Are the two distributions equal?
  - If the two distributions which we are comparing are exactly equal, then the points on the Q-Q plot will perfectly lie on a straight-line y = x.

```
{r}
df_qq %>%
  pivot_longer(cols = c("y1", "y2")) %>%
  ggplot(aes(x= value, col = name )) +geom_density()
```

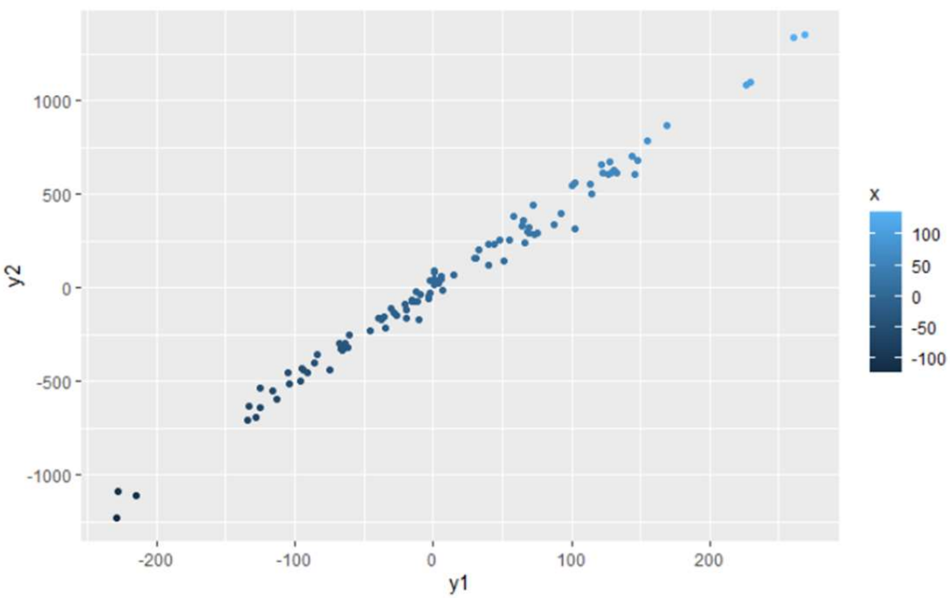


Let's test if this hold… So, we simulate some data.

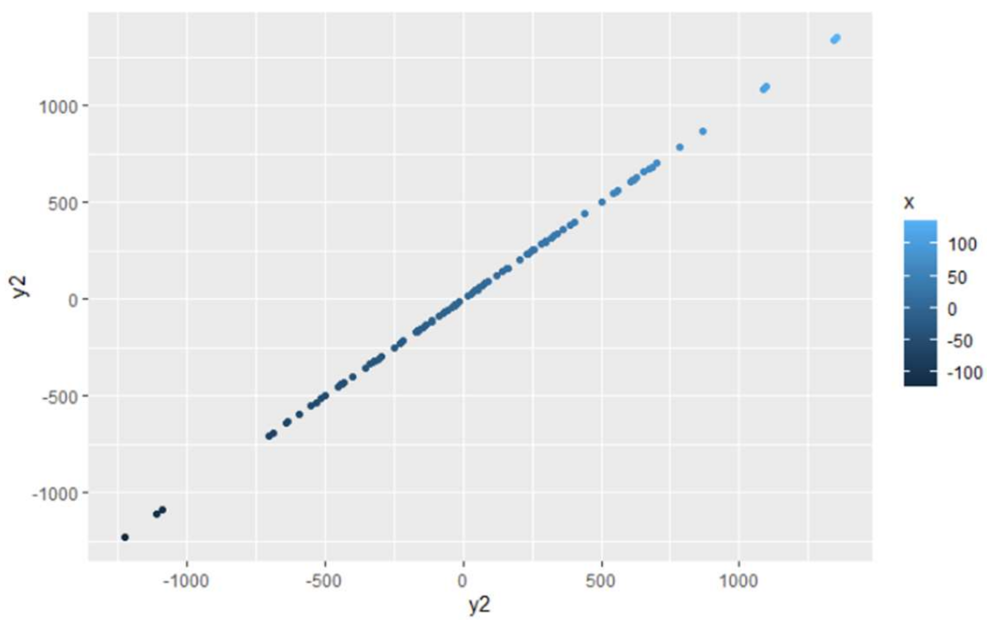
```
{r}
x <- rnorm(1e2,0,50)

df_qq <- data.frame(x = x, y1 = rnorm(1e2, mean = x * 2, sd = 10),  y2 = rnorm(1e2, mean = x*10, sd = 2))
```

Q-Q PLOTS



1000
500
0
-500
-1000
y2
-200
-100
0
100
200
y1
x
100
50
0
-50
-100



1000
500
0
-500
-1000
y2
-1000
-500
0
500
1000
y2
x
100
50
0
-50
-100

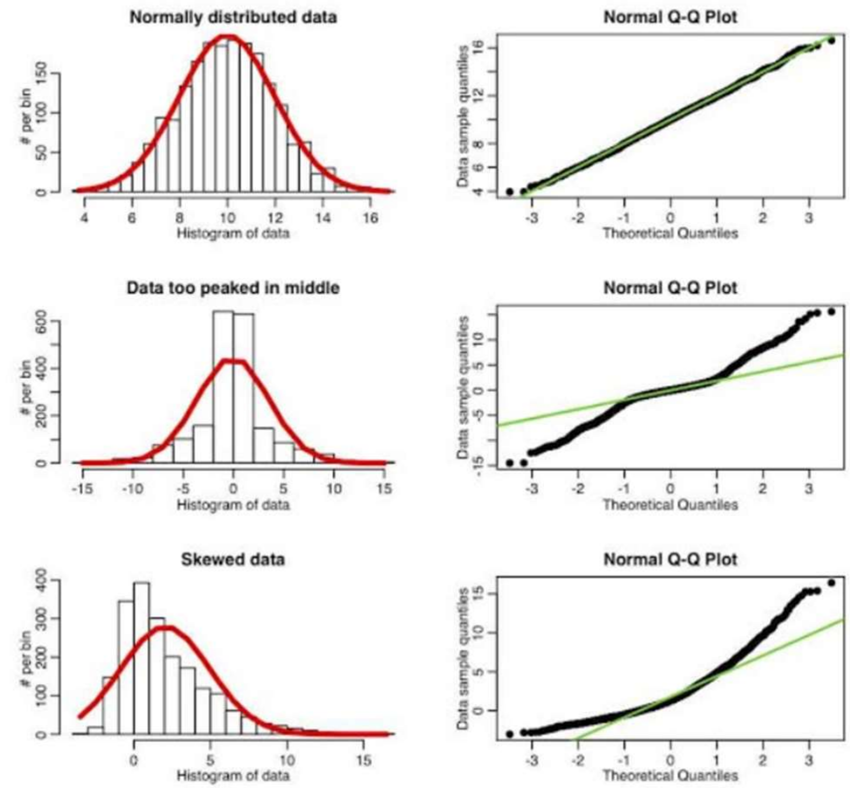
Q-Q PLOTS

We use it to investigate the distribution of a dataset.

- Normal distributions.
- Because we know certain percentiles correspond to specific intervals measured in standard deviations from the mean/median (same in a gaussian distributions).
- So, we our standardize variable.
    - Having our perfect theoretical gaussian distribution quantiles on the x-axis and y-axis has our standardized variable.

<u>**The question is:**</u>

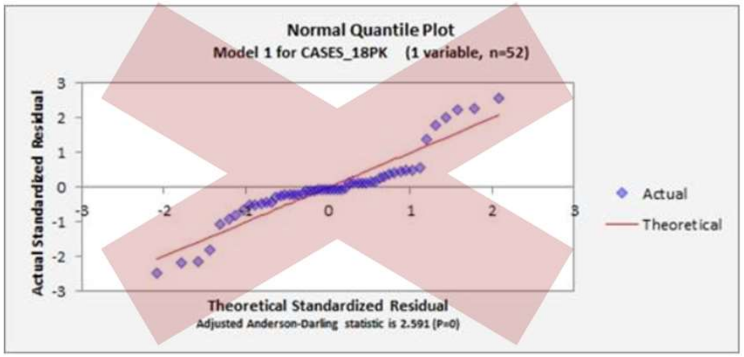
- Does our standardized variable match a perfect gaussian distribution quantile?



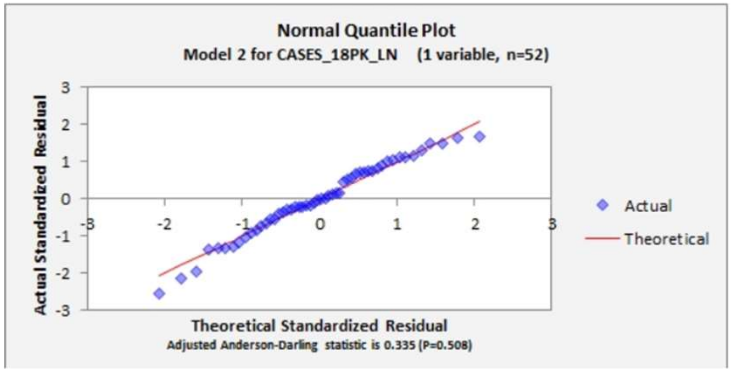
NORMALITY OF ERRORS.

Is our residuals normally distributed?

- Check with a Q-Q plot.



...and here is an example of a good-looking one (a linear pattern with P=0.5 for the A-D stat, indicating no significant departure from nor



COMPARING DATA TO REPLICATIONS FROM A FITTED MODEL

Data:

```
#Data
n <- 100
test_score <- rnorm(n, 15, sd = 5)
IQ <- rnorm(n, mean = test_score * 10, sd = 15)
data <- data.frame(IQ, test_score)
```

Model:

```
#Model
fit_test <- stan_glm(IQ ~ test_score, data = data, refresh = 0, prior = normal(location = 0,
scale = 10, autoscale =  T))

#Check Model
plot(fit_test)
prior_summary(fit_test)
```

COMPARING DATA TO REPLICATIONS FROM A FITTED MODEL

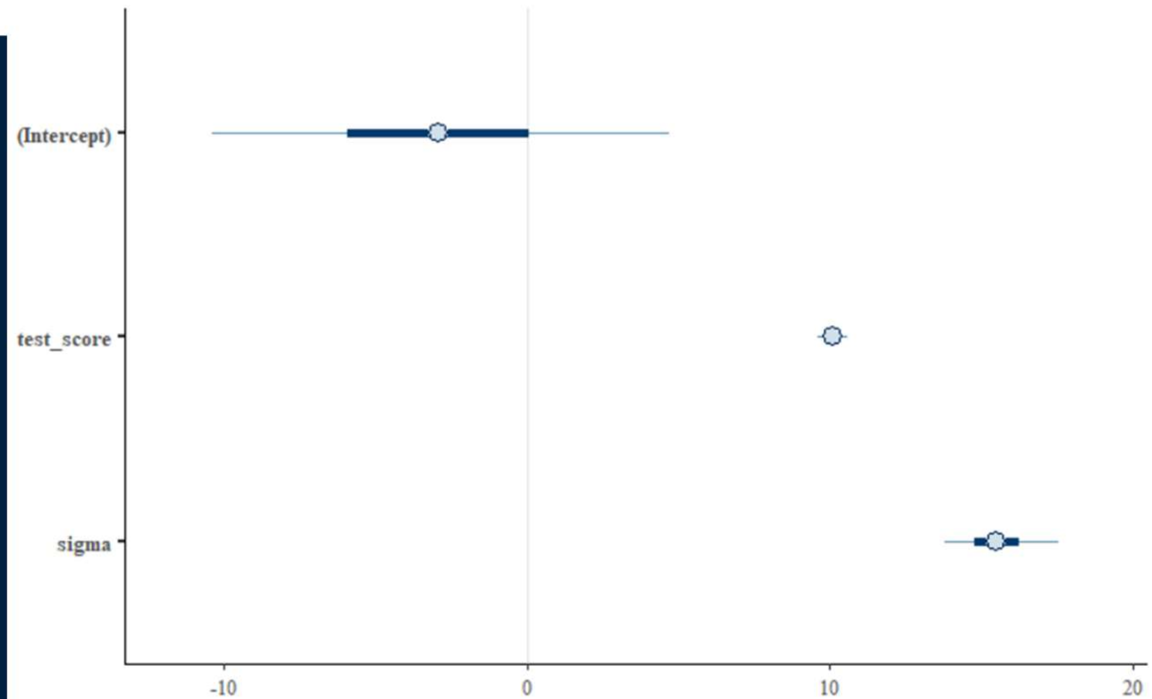
```
prior_summary(fit_test)
```

```
Priors for model 'fit_test'
------
Intercept (after predictors centered)
  Specified prior:
    ~ normal(location = 160, scale = 2.5)
  Adjusted prior:
    ~ normal(location = 160, scale = 109)

Coefficients
  Specified prior:
    ~ normal(location = 0, scale = 1)
  Adjusted prior:
    ~ normal(location = 0, scale = 10)

Auxiliary (sigma)
  Specified prior:
    ~ exponential(rate = 1)
  Adjusted prior:
    ~ exponential(rate = 0.023)
------
See help('prior_summary.stanreg') for more details
```

```
plot(fit_test)
```



COMPARING DATA TO REPLICATIONS FROM A FITTED MODEL

We remember that stan_glm() uses sampling.

```
head(as.matrix(fit_test))
```

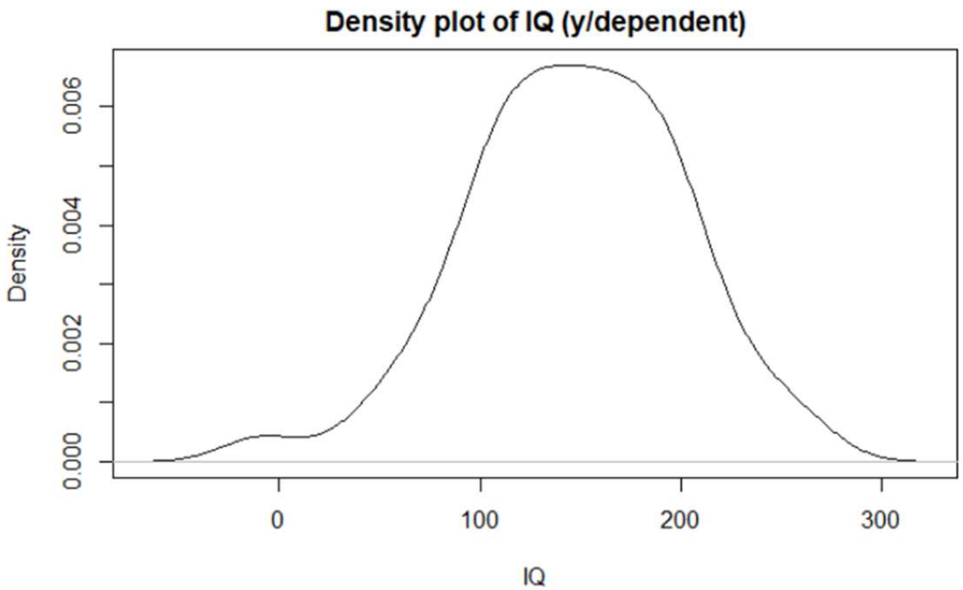
```
          parameters
iterations (Intercept) test_score    sigma
      [1,]   1.6591451   9.825943 15.80277
      [2,]  -4.7249498  10.161900 15.52715
      [3,]  -6.7900433  10.276070 15.50426
      [4,]  -0.1591298   9.978776 16.30445
      [5,]  -3.6230020  10.053522 15.71305
      [6,]   2.6706013   9.673399 14.83182
```

We want to check how well our model is doing...

- Could use AIC, BIC, R2, Bayesian R2, -log-likelihood.
- We COULD also check the posterior predictive distribution.

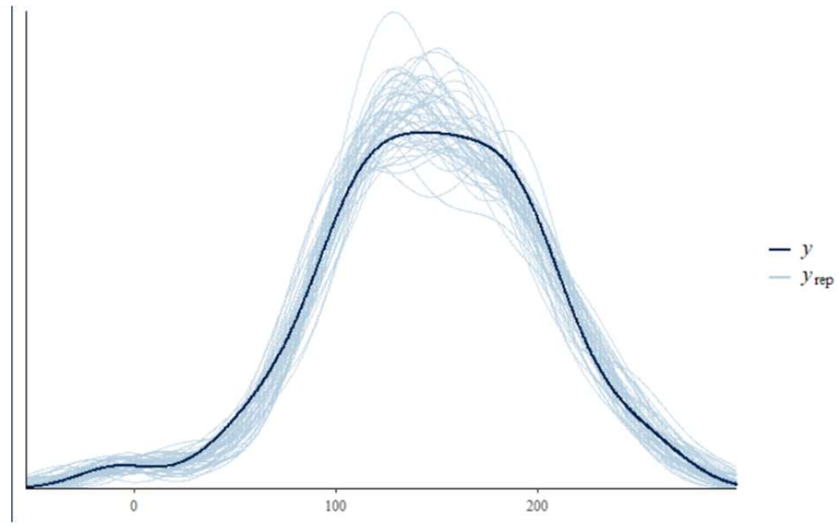
COMPARING DATA TO REPLICATIONS FROM A FITTED MODEL

Our observed IQ distribution.



Observed and predicted.

- Using different θ vectors and the original X matrix.




```
head(as.matrix(fit_test))
```

```
          parameters
 iterations (Intercept) test_score    sigma
      [1,]   1.6591451   9.825943 15.80277
      [2,]  -4.7249498  10.161900 15.52715
      [3,]  -6.7900433  10.276070 15.50426
      [4,]  -0.1591298   9.978776 16.30445
      [5,]  -3.6230020  10.053522 15.71305
      [6,]   2.6706013   9.673399 14.83182
```

A BAD EXAMPLE

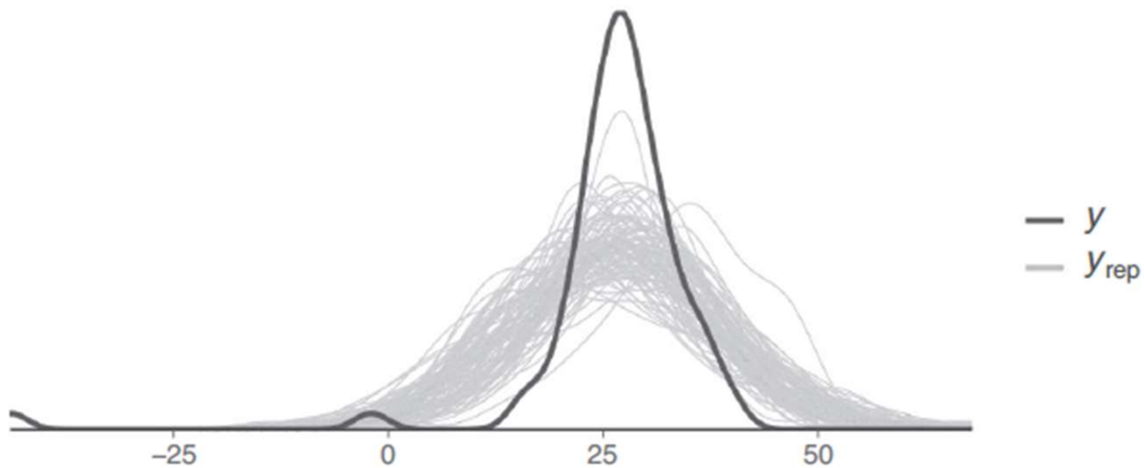
- All the replication fits the observed y distribution poorly.

- We've observed y-values of -30. While non of the posterior draws predict anything below -25.
- Further it's slightly too flat.

- Not necessarily a success criteria that they're alike. But gives intuition of your model restrictions.

- There is something in the data that our model isn't catching

- **Assumption:** That the distribution is representative and we've enough samples.



EXERCISES:

- 10.1
  - In addition: With and without interaction do the following
  - Also use the posterior predictive check to see if the predicted density fits the acutal density distribution y.
- 10.9
- 11.5
- 11.9
- 11.3

AARHUS
UNIVERSITY