ABI Summer 2021 Introduction to Statistics

Guest Session 4

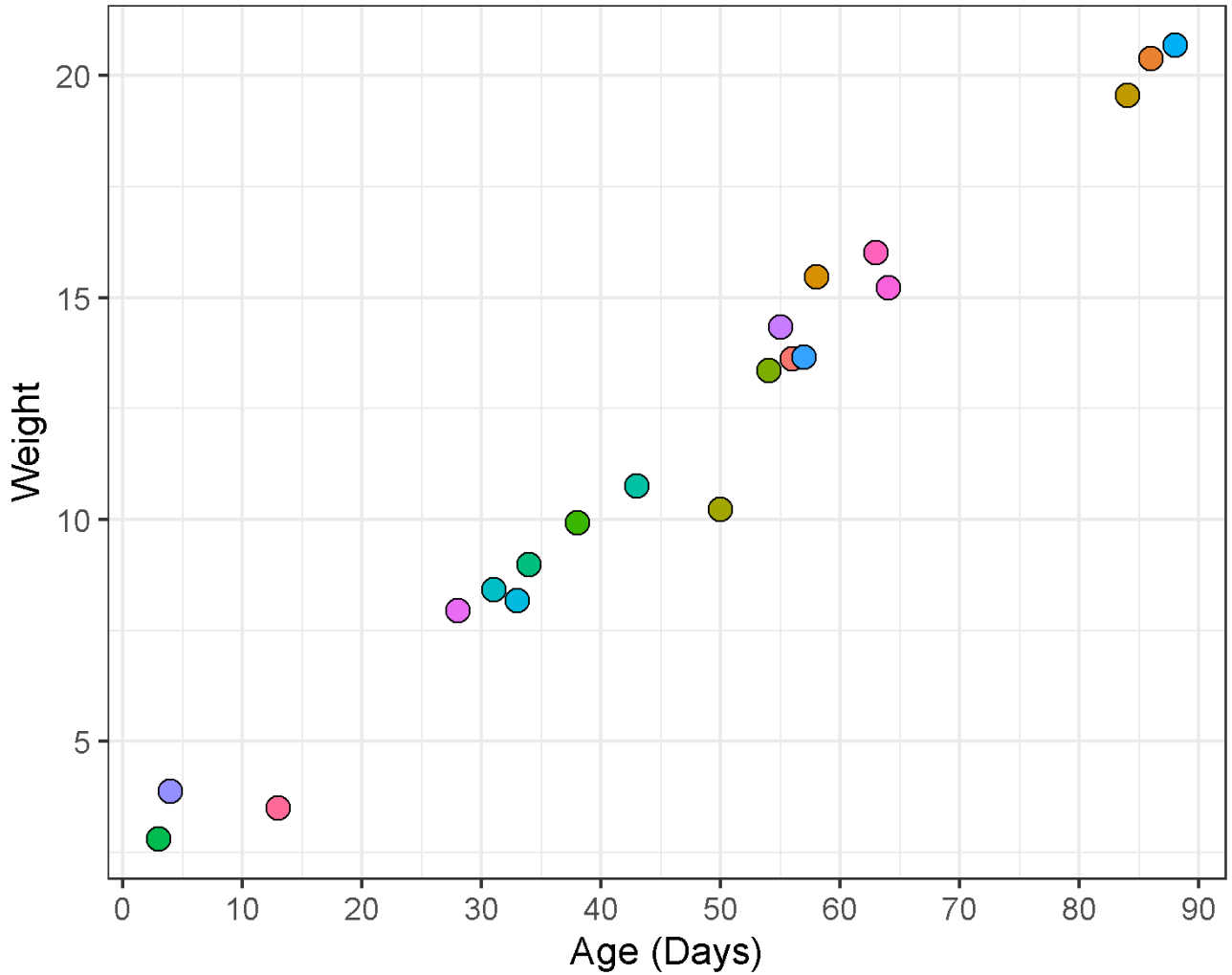
Javier Cabrera*, Volha Tryputsen** & Davit Sargsyan**

July 15, 2021

(*)Rutgers, the State University of New Jersey
(**)Janssen Pharmaceutical Research and Development (J&J)

Body Weight vs. Age

- Twenty (20) mice between the age of 0 and 90 days were weighted
- What is the relationship between age and weight?

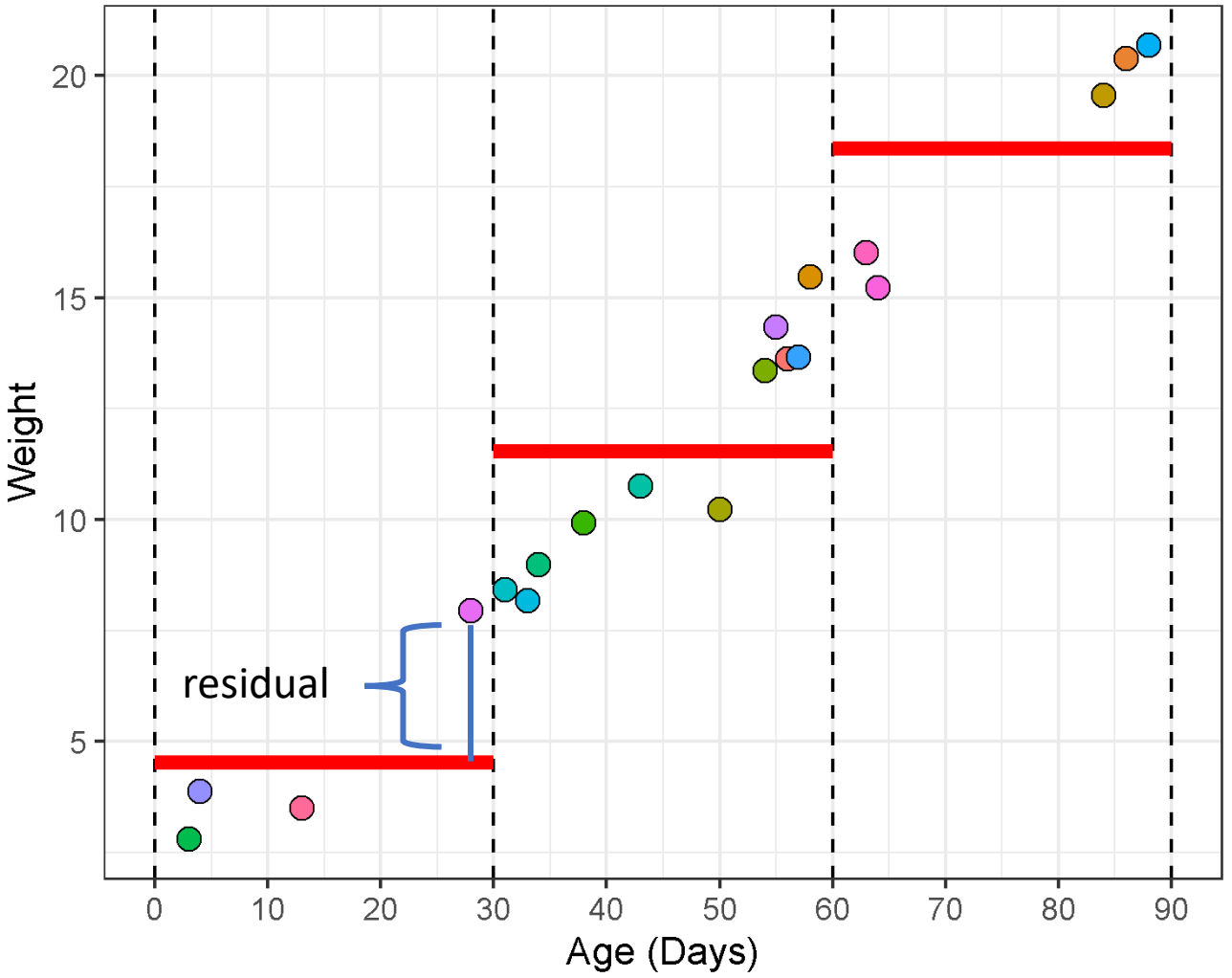


Average Weight by Age

- We can bin the mice by age, e.g., less than 1, 2 and 3 months old.
- Calculate the mean weight by age in months:

Age (Months)	Mean	SEM
1	4.53	1.16
2	11.54	0.79
3	18.37	1.15

- Mice weight is positively associated with age (older mice weight more)



Line of Best Fit

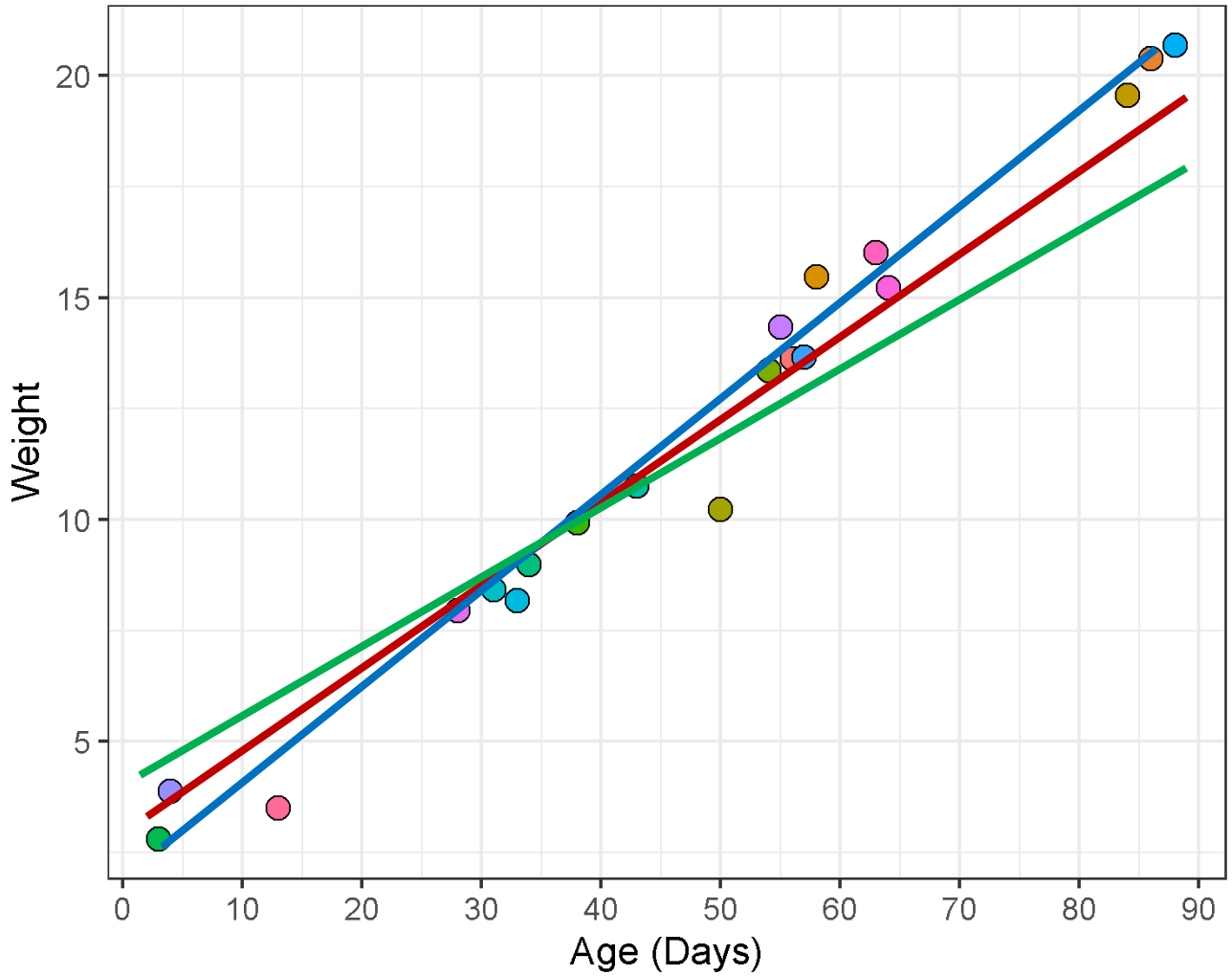
- Instead of binning the observation, lets try to find a relationship that we can describe with a line:

$$Y=a + bX$$

Where a is the y-intercept and b is the slope of the line

- There are many lines we can draw trough the data

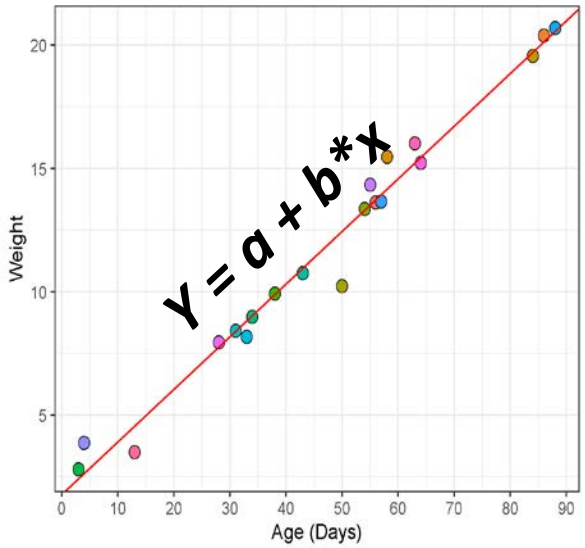
- How do we find the best one? What does it mean – a line of best fit?



Linear Regression

- We can try to find a line that will minimize the sum of squares of residuals by fitting one that looks good and change intercept (***a***) and slope (***b***) until we cannot do any better

- Instead, we can solve for a and b: the method of **least squares.**

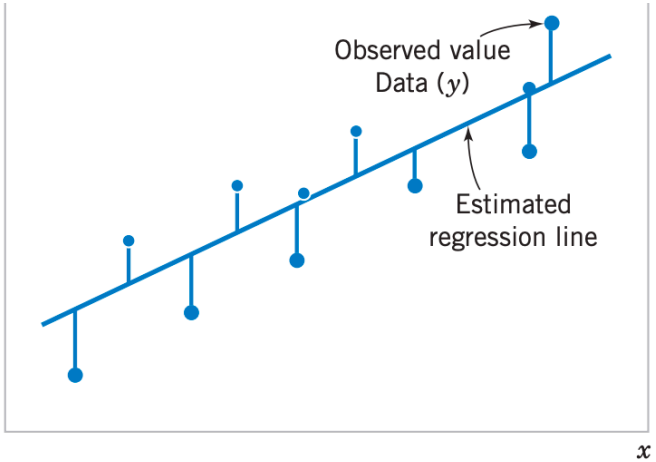


R output (function ***lm***)

```
Call:
lm(formula = Weight ~ Days, data = dtg)

Residuals:
     Min       1Q   Median       3Q      Max
-2.23440 -0.21449  0.01892  0.25992  1.29533

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.787496   0.392868    4.55 0.000248 ***
Days        0.213480   0.007429   28.73  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7988 on 18 degrees of freedom
Multiple R-squared:  0.9787,    Adjusted R-squared:  0.9775
F-statistic: 825.7 on 1 and 18 DF,  p-value: < 2.2e-16

(Intercept)        Days
  1.7874963   0.2134798
```

Method of least squares

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, \ldots, n$$



To estimate (β_0,β_1) , we find values that minimize squared error:

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The least squares estimators of β_0 and β_1, say, $\hat{\beta}_0$ and $\hat{\beta}_1$, must satisfy

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

Least square normal equations

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

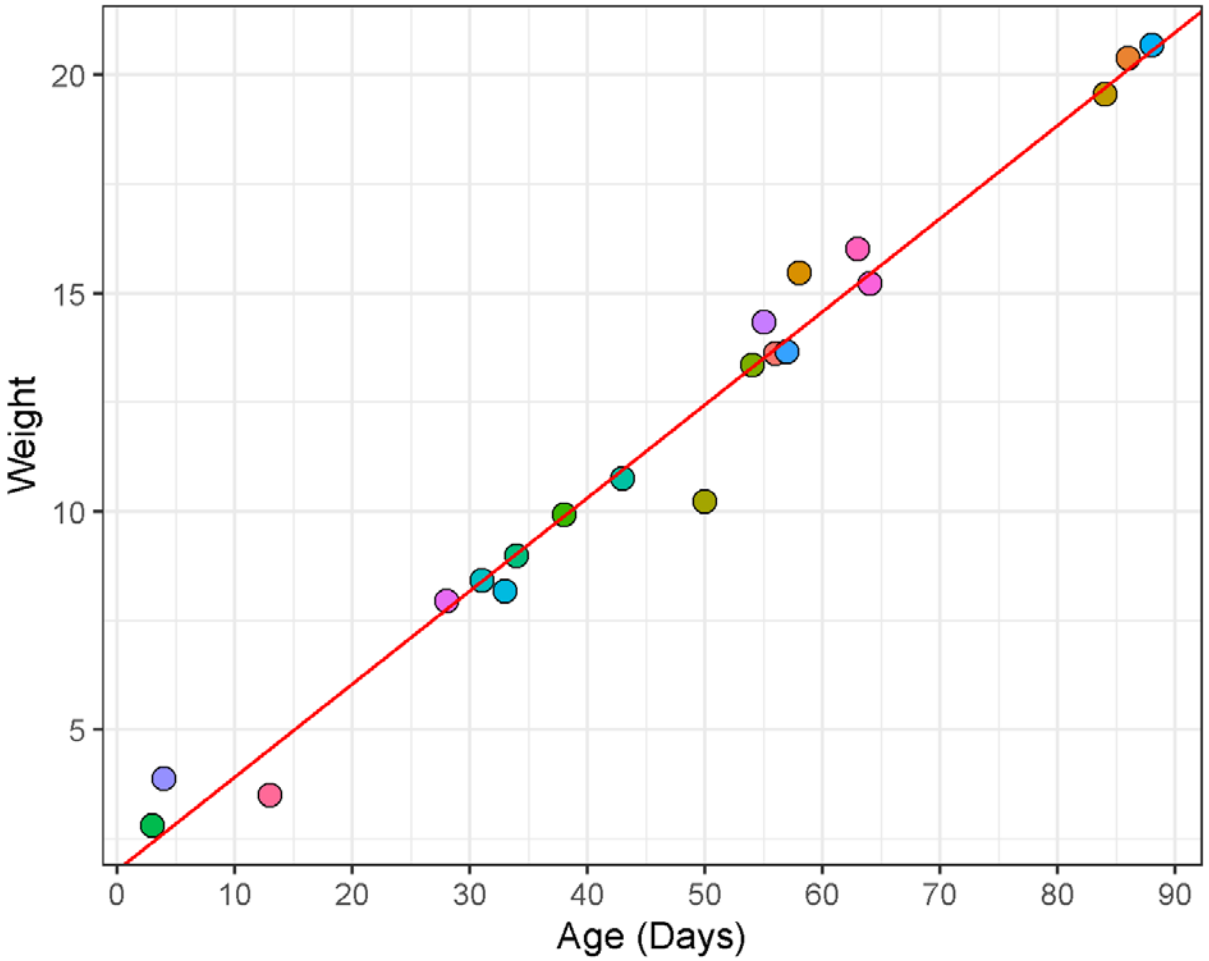
The least squares estimates of the intercept and slope in the simple linear regression model are:

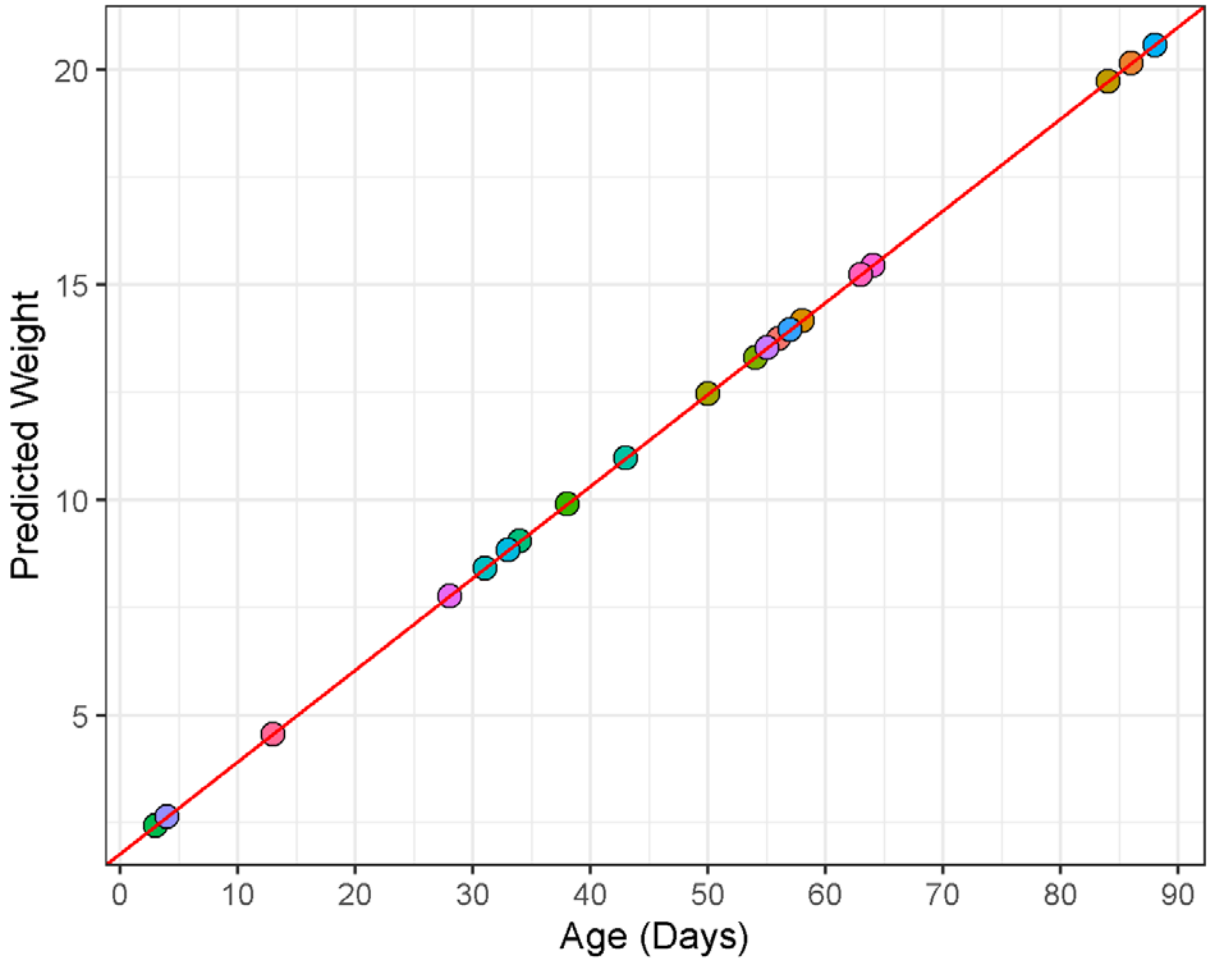
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

where $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ and $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$.

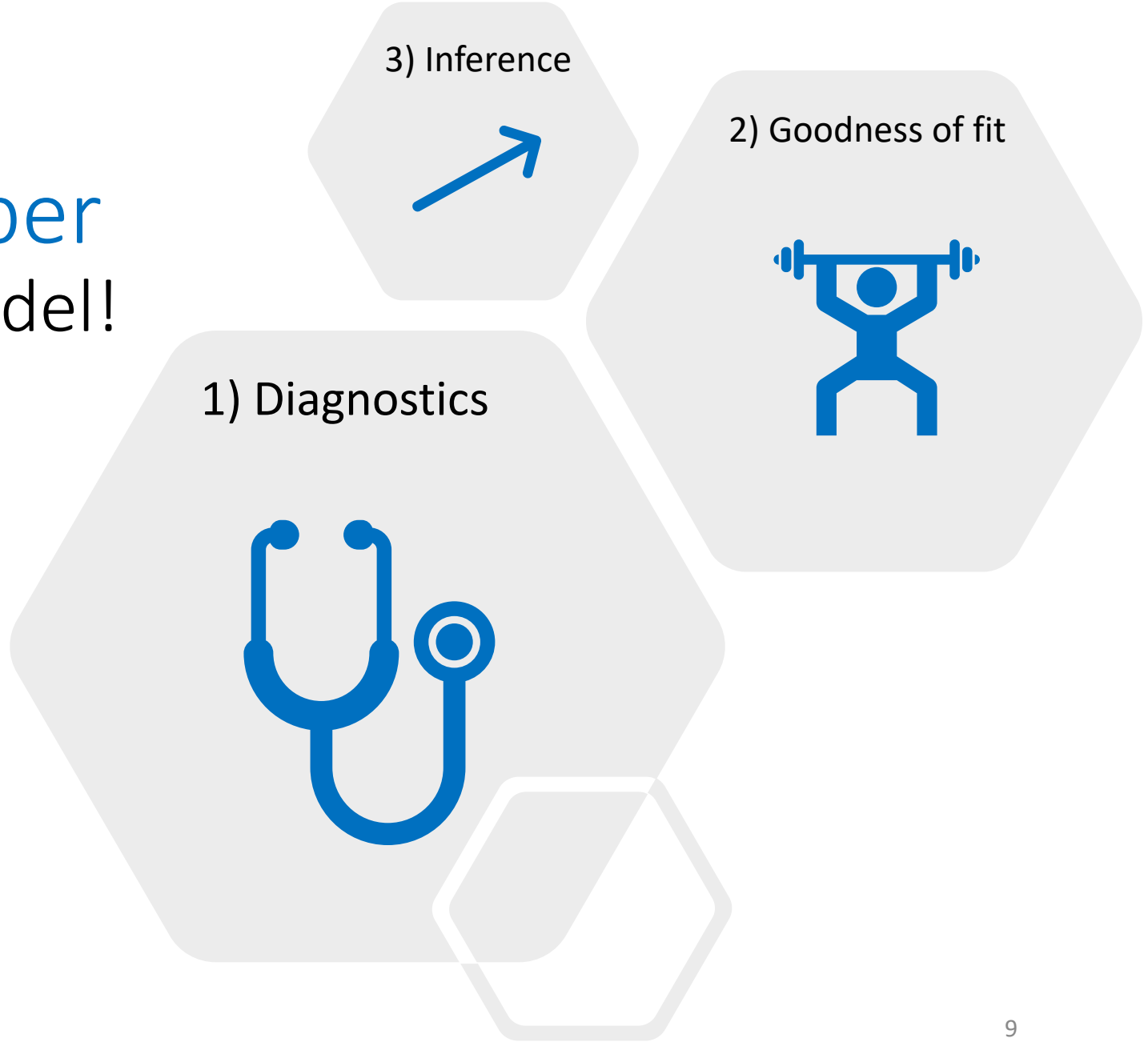
Predicting Observations



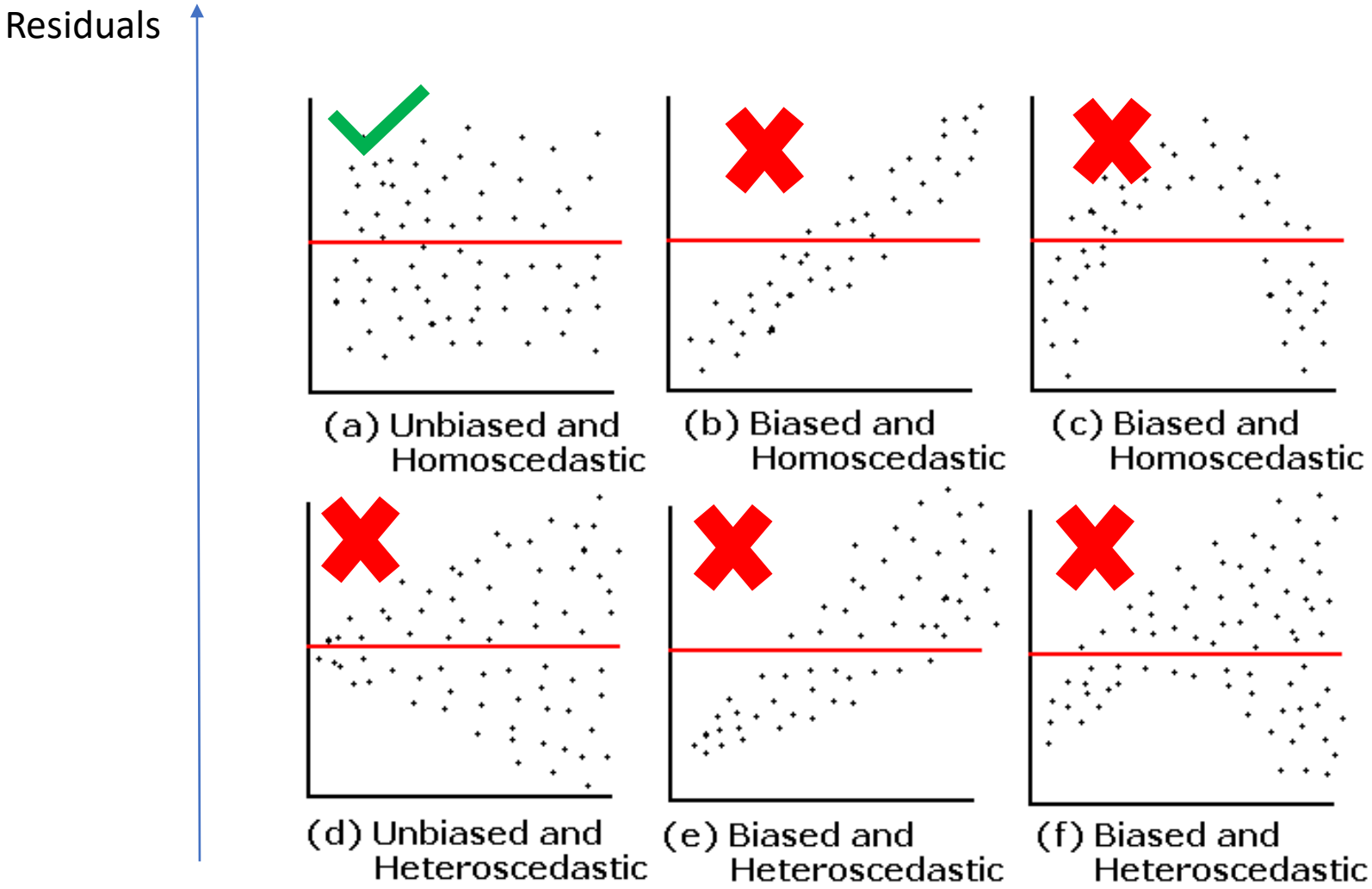


3 things to remember

when fitting a linear model!

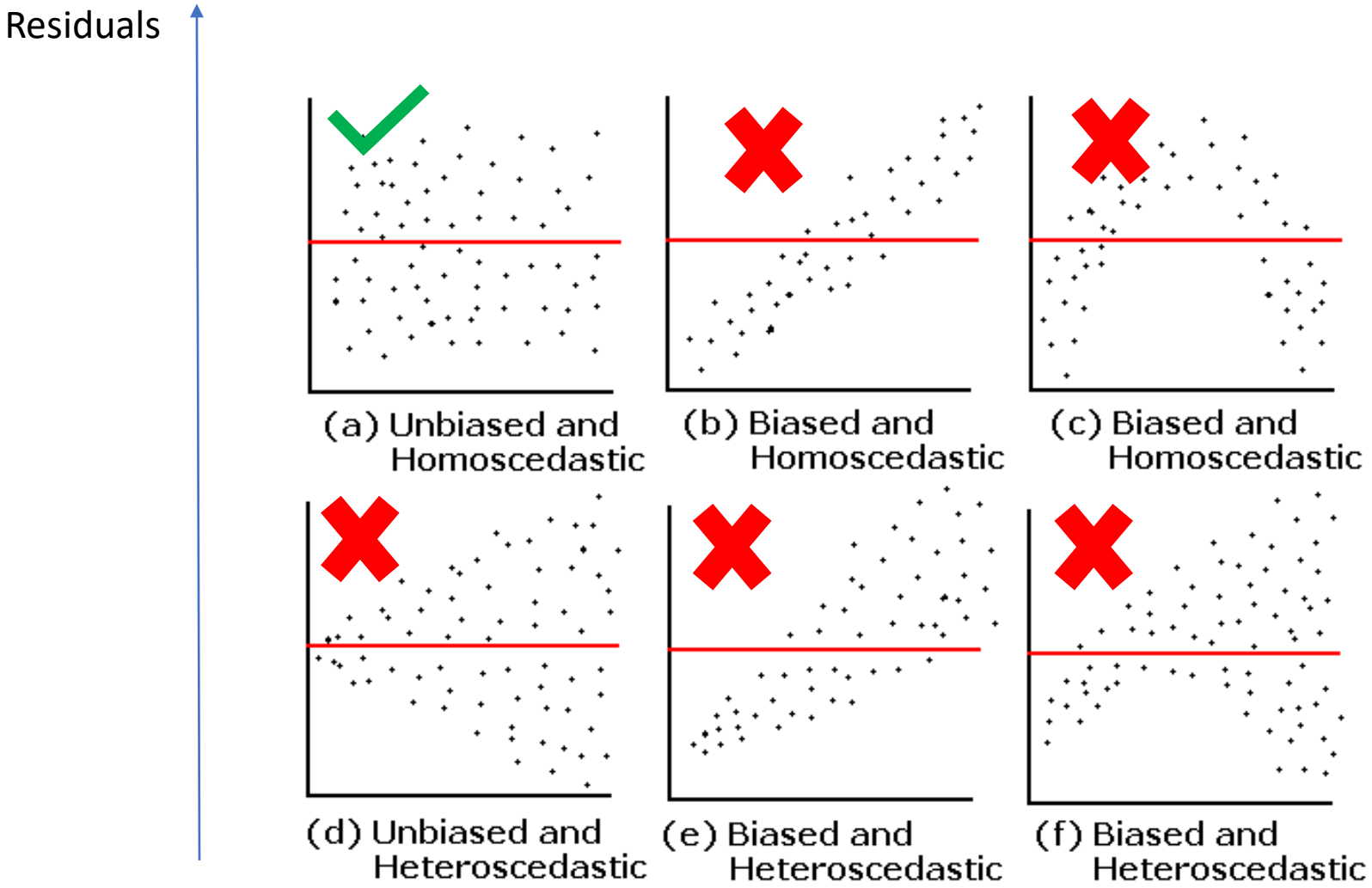


(1) Model Diagnostics: residuals vs fitted



! If residuals are scattered randomly around 0 with uniform variation, it indicates that the data fit a linear mode!

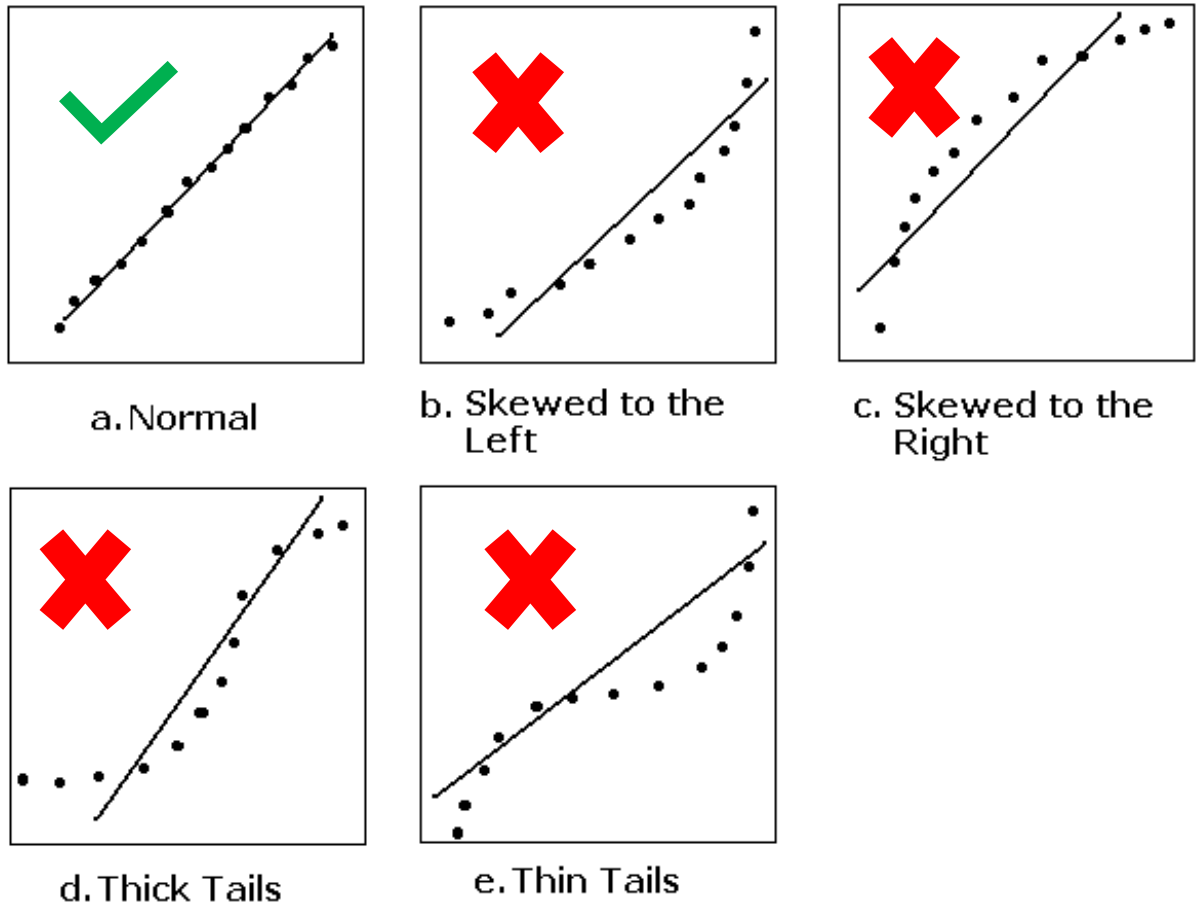
(1) Model Diagnostics: residuals vs fitted



"Don't stop building your model until residuals show no pattern"
-John Tukey

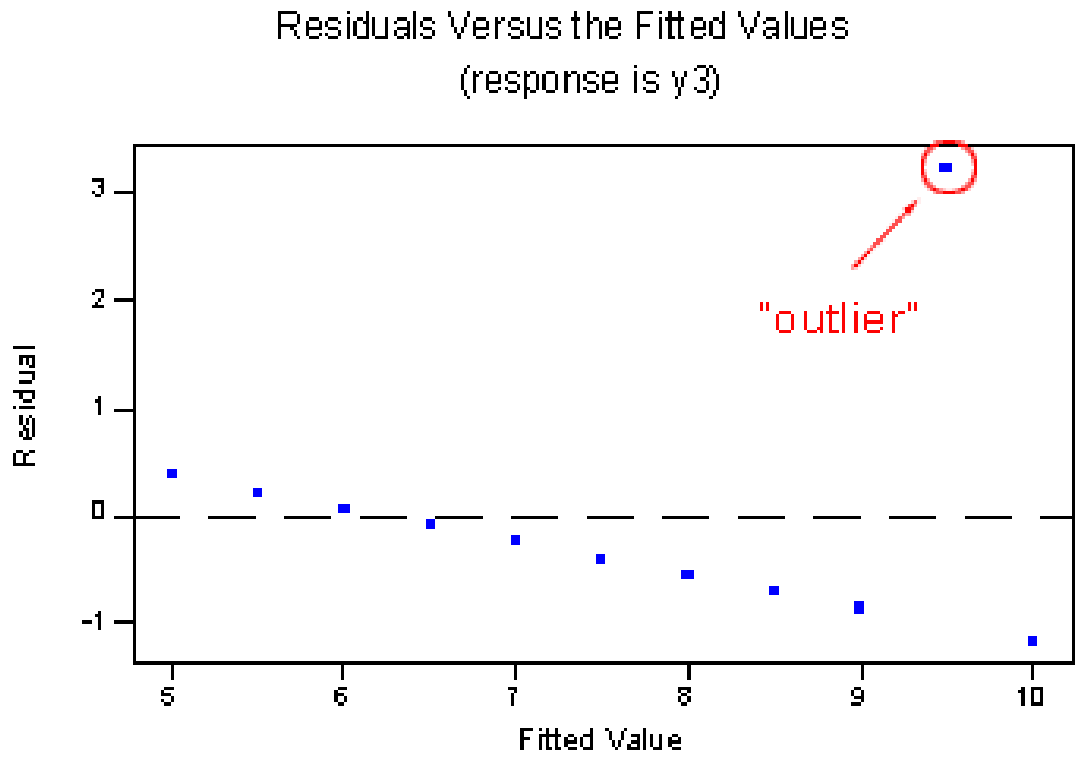
(1) Model Diagnostics: normality of residuals

qq-plot of residuals



! The line on the plot is straight, supporting the assumption of normally distributed residuals!

(1) Model Diagnostics: outliers



An **outlier** is an observation that has a large **residual**. In other words, **the** observed value for **the** point is very different from that predicted by **the regression model**.

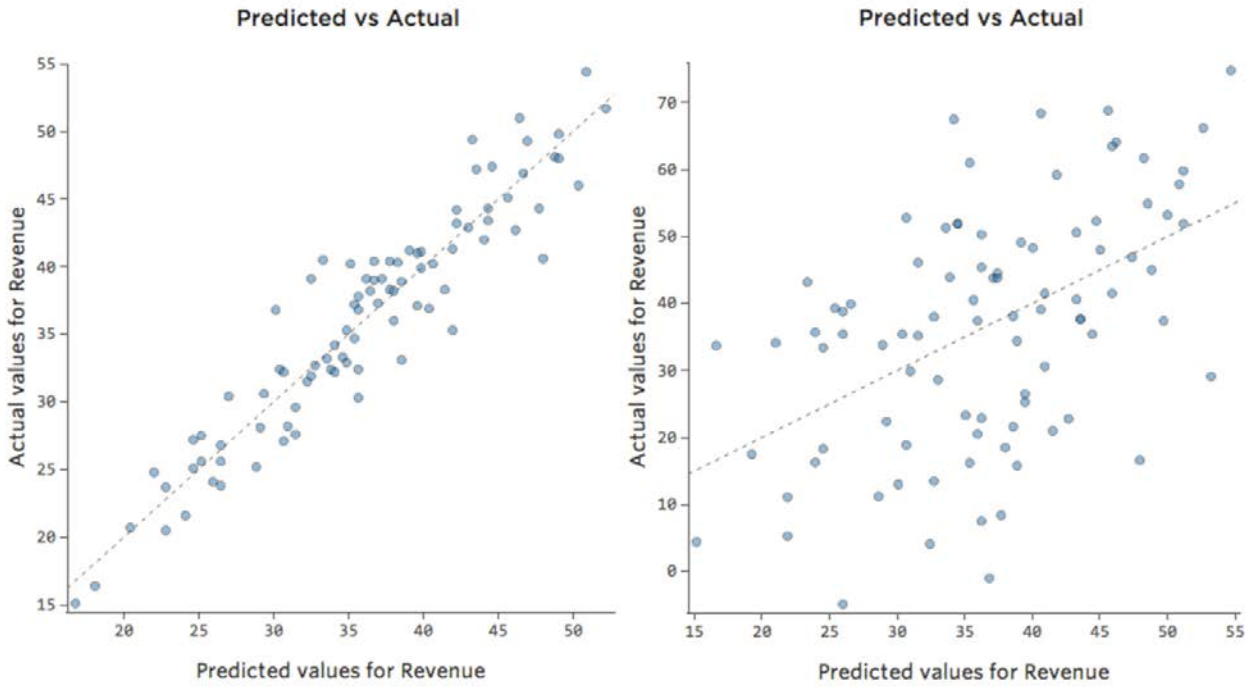
Q: What to do when data has outliers?

A: Use **robust regression**!

(2) Model Performance: Goodness of Fit

- How well you can predict a future dataset?
- R-squared is displayed in the summary(lm(y~x))
- Look at the plot of predicted values vs. observed
- R-squared = correlation(predicted vs. observed)^2

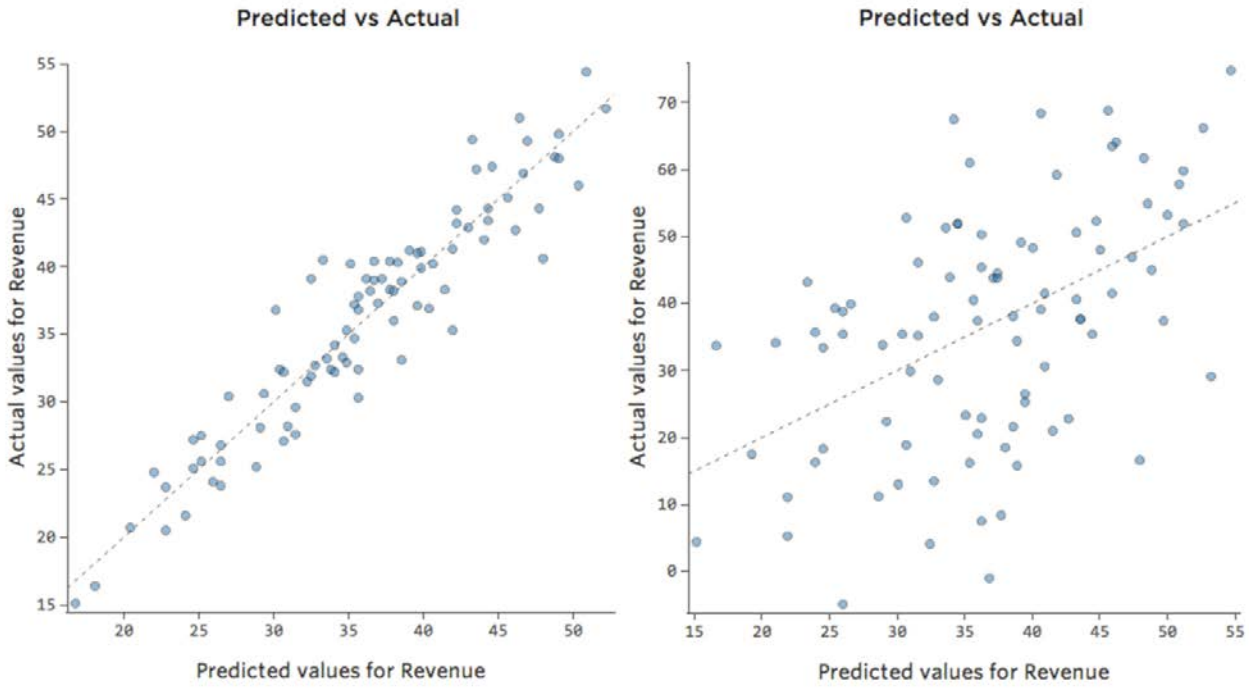
- *NOTE: even if R-squared is low but the slope is significant, model can still be very useful to establish a relationship*



(2) Model Performance: Goodness of Fit

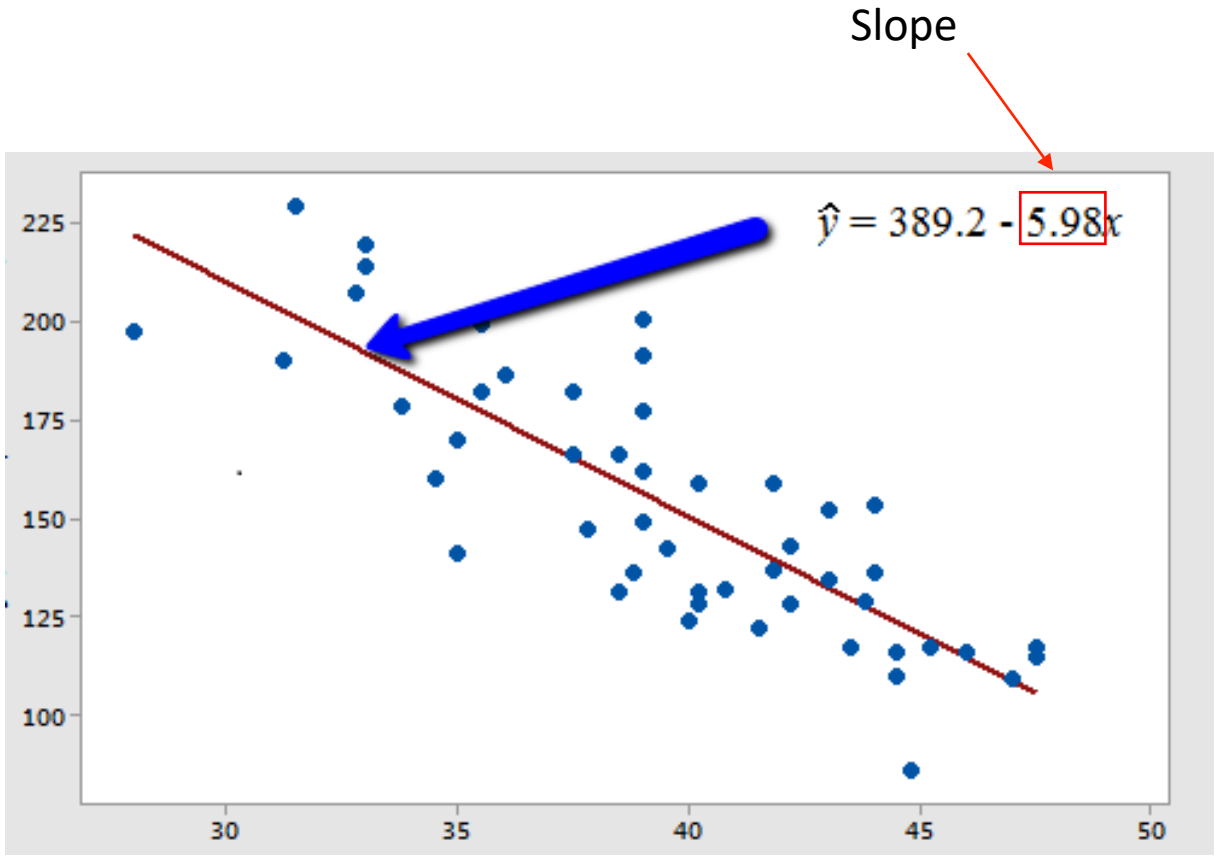
- How well you can predict a future dataset?
- R-squared is displayed in the summary(lm(y~x))
- Look at the plot of predicted values vs. observed
- R-squared = correlation(predicted vs. observed)^2

- *NOTE: even if R-squared is low but the slope is significant, model can still be very useful to establish a relationship*



"All models are wrong, but some are useful"
- George Box

(3) Inference: Estimating the Slope

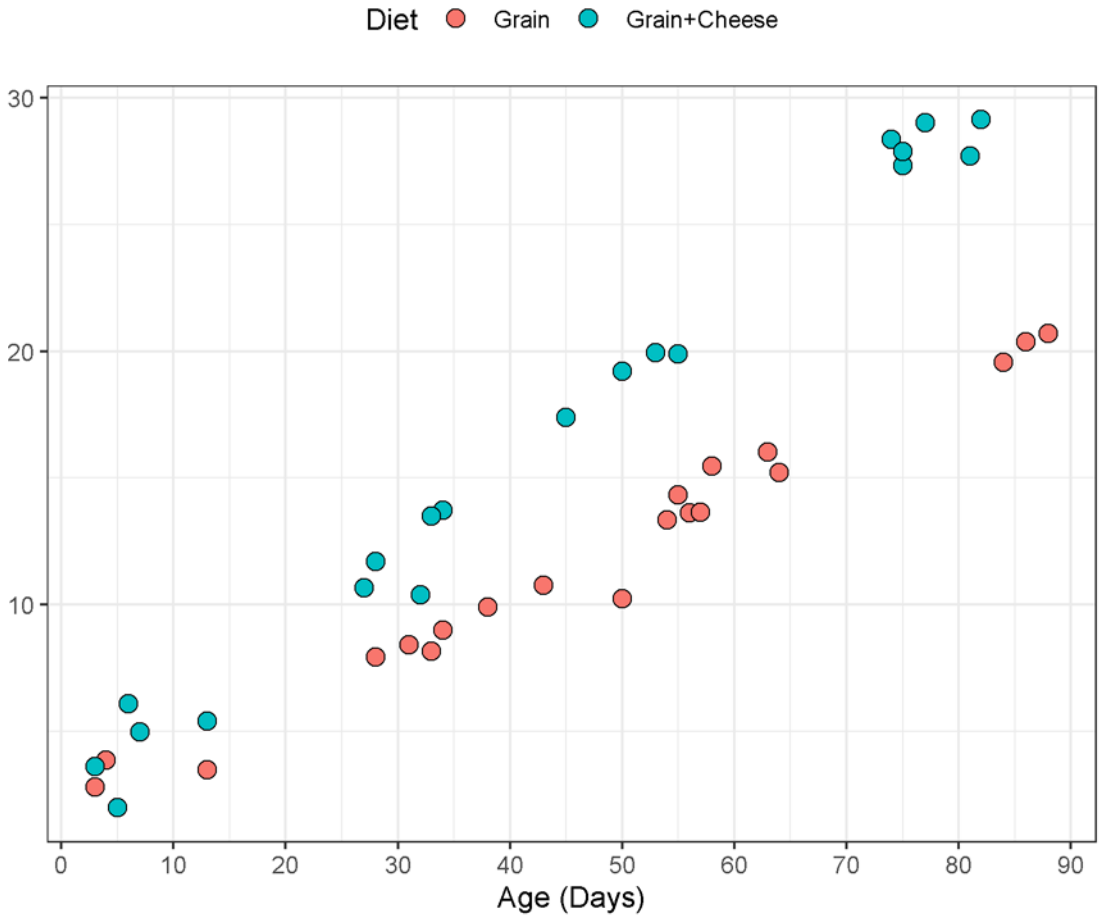


- **Slope of a linear regression line** tells us how much change in y-variable is caused by a unit change in x-variable.
- Q: Is it statistically significantly different than "zero"?
- A: For that, we can test the hypothesis that the regression slope parameter β is equal to zero. $H_0: \beta_1 = 0$ vs. $H_0: \beta_1 \neq 0$
- We calculate $t = b_1 / SE_{b1}$ to find a p-value for the t-test of significance of the slope.

More than one predictors

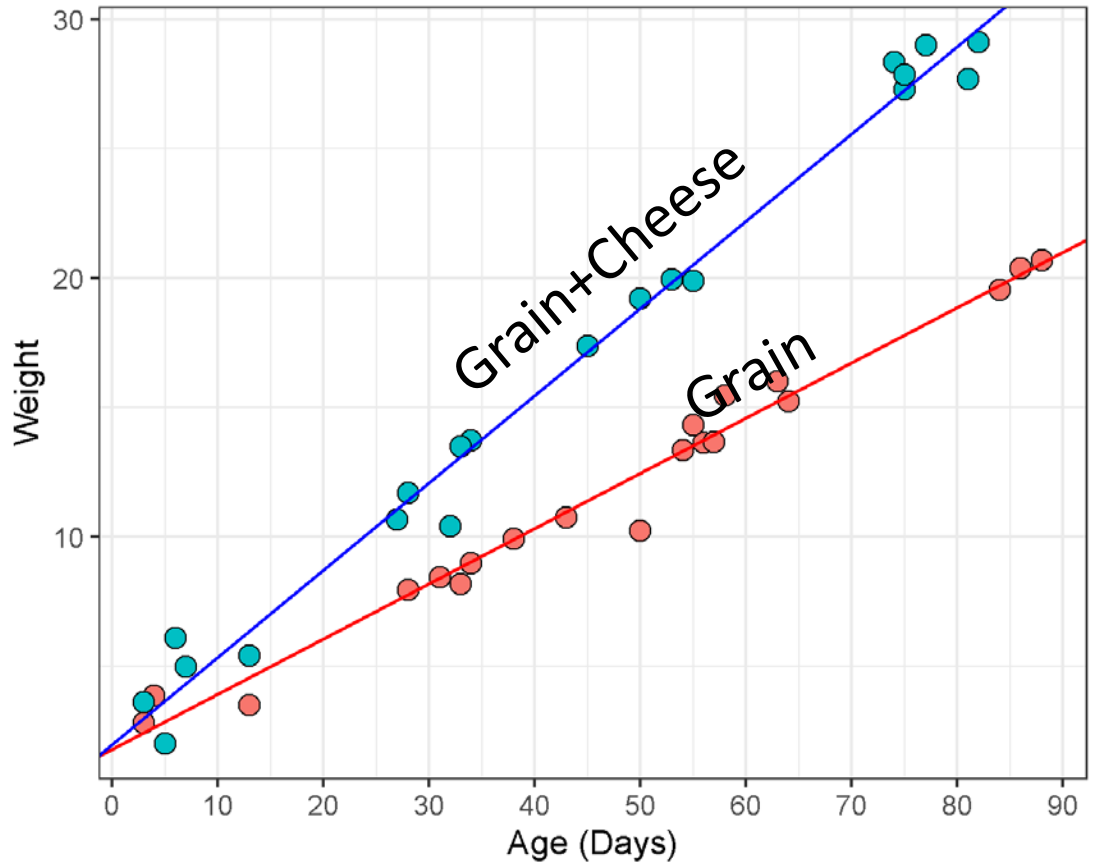
- Now, let's add another variable to explain the weight difference: Diet

```
Call:
lm(formula = Weight ~ Days + Diet, data = dt1)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7020 -1.2654  0.0955  0.9248  4.2371

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.49437    0.69508  -2.150   0.0382 *
Days              0.28316    0.01169  24.229  < 2e-16 ***
DietGrain+Cheese  5.78058    0.60238   9.596 1.39e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.898 on 37 degrees of freedom
Multiple R-squared:  0.9457,    Adjusted R-squared:  0.9428
F-statistic: 322.2 on 2 and 37 DF,  p-value: < 2.2e-16
```



Interaction

```
Call:
lm(formula = Weight ~ Days * Diet, data = dt1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.36887 -0.32586  0.05597  0.43711  2.11267

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.787496   0.473711   3.773 0.000581 ***
Days                0.213480   0.008958  23.831  < 2e-16 ***
DietGrain+Cheese    0.174880   0.620641   0.282 0.779731
Days:DietGrain+Cheese 0.124037   0.011952  10.378 2.28e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9631 on 36 degrees of freedom
Multiple R-squared:  0.9864,    Adjusted R-squared:  0.9853
F-statistic: 870.2 on 3 and 36 DF,  p-value: < 2.2e-16
```



LINEAR MODEL FOR TWO FACTORS

Model: Data = Tot Mean + Factor 1 Effect + Factor 2 Eff + Error

LINEAR MODEL FOR TWO WAY TABLE of gene expressions:

This is a typical dataset where we have a response and two factors:
Factor 1 is the sample and Factor 2 is the probe and we want to estimate the sample effect .
Example: Observe the 11 probes for one gene in 6 samples

	C1	C2	C3	T1	T2	T3
Probe	1521b99	1532b99	2353b99	1521a99	1532a99	2353a99
1	137.08	165.92	112.41	168.68	83.05	103.02
2	603.08	605.42	420.91	681.18	534.05	479.02
3	851.08	981.42	724.91	989.18	809.35	717.02
4	19.76	15.91	22.29	31.22	30.23	13.57
5	237.08	193.42	136.41	227.18	255.05	114.02
6	77.58	128.22	83.21	87.18	91.85	41.98
7	1212.58	1188.22	818.91	1279.18	1279.05	959.32
8	759.38	798.42	770.91	857.48	1175.05	868.52
9	84.38	110.42	109.71	122.18	112.35	63.07
10	41.59	35.74	40.71	43.18	34.12	22.32
11	158.58	140.42	135.91	162.48	172.05	93.02

	exprs	probe	sample
1	168.68	1	T1
2	681.18	2	T1
3	989.18	3	T1
4	31.22	4	T1
5	227.18	5	T1
6	87.18	6	T1
7	1279.18	7	T1
8	857.48	8	T1
9	122.18	9	T1
10	43.18	10	T1
11	162.48	11	T1
12	83.05	1	T2
13	534.05	2	T2
14	809.35	3	T2
15	30.23	4	T2
16	255.05	5	T2
17	91.85	6	T2
18	1279.05	7	T2
19	1175.05	8	T2
20	112.35	9	T2
21	34.12	10	T2
22	172.05	11	T2
23	103.02	1	T3
24	479.02	2	T3
25	717.02	3	T3
26	13.57	4	T3
27	114.02	5	T3
28	41.98	6	T3
...	...	...	...

LINEAR MODEL FOR TWO FACTORS

```
summary(lm(exprs~ probe + sample,data))
Residuals:
    Min      1Q  Median      3Q     Max
-237.84  -34.81   -9.40   41.93  260.47
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   135.49      38.92   3.482  0.00104 **
probe2        425.58      45.63   9.326 1.59e-12 ***
probe3        717.13      45.63  15.715  < 2e-16 ***
probe4       -106.20      45.63  -2.327  0.02405 *
probe5         65.50      45.63   1.435  0.15741
probe6        -43.36      45.63  -0.950  0.34663
probe7        994.52      45.63  21.794  < 2e-16 ***
probe8        743.27      45.63  16.288  < 2e-16 ***
probe9        -28.01      45.63  -0.614  0.54215
probe10       -92.08      45.63  -2.018  0.04899 *
probe11        15.38      45.63   0.337  0.73745
sampleC2       16.49      33.70   0.489  0.62684
sampleC3      -73.26      33.70  -2.174  0.03448 *
sampleT1       42.45      33.70   1.260  0.21368
sampleT2       35.82      33.70   1.063  0.29295
sampleT3      -64.30      33.70  -1.908  0.06216 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 79.04 on 50 degrees of freedom
Multiple R-squared:  0.969,         Adjusted R-squared:  0.9597
F-statistic: 104.3 on 15 and 50 DF,  p-value: < 2.2e-16
```

	C1	C2	C3	T1	T2	T3
Probe	1521b99	1532b99	2353b99	1521a99	1532a99	2353a99
1	137.08	165.92	112.41	168.68	83.05	103.02
2	603.08	605.42	420.91	681.18	534.05	479.02
3	851.08	981.42	724.91	989.18	809.35	717.02
4	19.76	15.91	22.29	31.22	30.23	13.57
5	237.08	193.42	136.41	227.18	255.05	114.02
6	77.58	128.22	83.21	87.18	91.85	41.98
7	1212.58	1188.22	818.91	1279.18	1279.05	959.32
8	759.38	798.42	770.91	857.48	1175.05	868.52
9	84.38	110.42	109.71	122.18	112.35	63.07
10	41.59	35.74	40.71	43.18	34.12	22.32
11	158.58	140.42	135.91	162.48	172.05	93.02

Fitted model: Data = Total Sample Effect + Probe Effect + Residual
= (Intercept + sample)+ Probe Effect + Residual

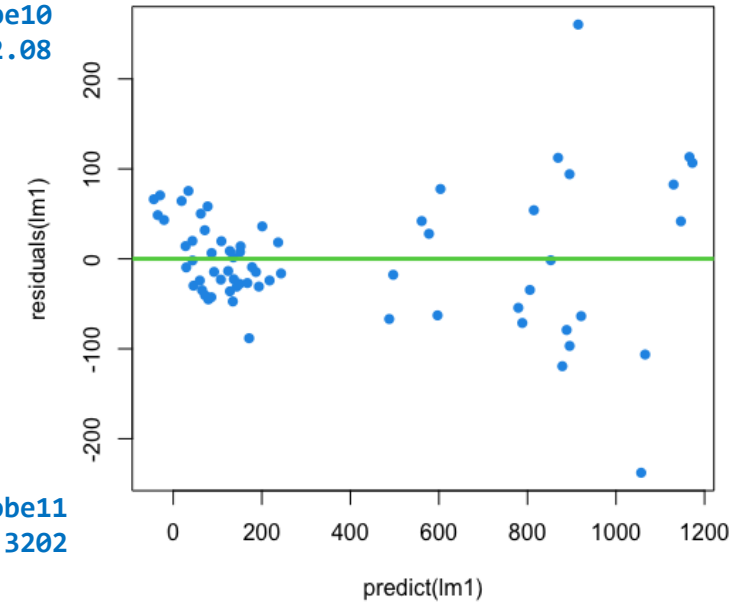
Condition: probe effects need to add up to zero

```
(Intercept)      probe2      probe3      probe4      probe5      probe6      probe7      probe8      probe9     probe10
     135.49      425.58      717.13     -106.20       65.50      -43.36      994.52      743.27      -28.01      -92.08
    probe11    sampleC2    sampleC3    sampleT1    sampleT2    sampleT3
      15.38       16.49      -73.26       42.45       35.82      -64.30
```

LINEAR MODEL FOR TWO FACTORS

Fitted model: Data = Total Sample Effect + Probe Effect + Residual
= (Intercept + sample)+ Probe Effect + Residual
Condition: probe effects need to add up to zero

```
(Intercept)      probe2      probe3      probe4      probe5      probe6   probe7   probe8   probe9   probe10
     135.49      425.58      717.13     -106.20       65.50      -43.36   994.52   743.27   -28.01    -92.08
    probe11    sampleC2    sampleC3    sampleT1    sampleT2    sampleT3
      15.38       16.49       -73.26       42.45       35.82      -64.30

sampleExpresLS = b[1] + c(sampleC1=0,b[12:16])
probeEffects = c(probe1=0,b[2:11])
mean(probeEffects)   [1] 244.7035

(sampleExpresLS =sampleExpresLS + mean(probeEffects))
sampleC1 sampleC2 sampleC3 sampleT1 sampleT2 sampleT3
380.1973 396.6845 306.9355 422.6473 416.0182 315.8982

(probeEffects =probeEffects - mean(probeEffects))
   probe1    probe2    probe3    probe4    probe5    probe6    probe7    probe8    probe9   probe10   probe11
-244.7035  180.8798  472.4298 -350.9002 -179.2035 -288.0602  749.8132  498.5632 -272.7118 -336.7868 -229.3202
  t.test(sampleExpresLS[1:3], sampleExpresLS[4:6])
t = -0.53359, df = 3.8138, p-value = 0.6232
mean of x mean of y
 361.2724  384.8545
```



Median Polish

Estimate the median rather than the mean

- Robust to outliers (Remember that Least Squares is not robust to outliers.)

Median Polish model : Data = Tot Median + Row Effect + Col Eff + Residual

Fitted model: Data = Total Sample Effect + Probe Effect + Residual

Total Sample Effect = Tot Median+ Col Effect

ALGORITHM FOR MEDIAN POLISH: Estimate Row Col and Tot effects using medians.

- Alternate iteration of removing the medians of rows and columns

- Continue iteration until reduction is less than a constant in sum of residual squares or absolute residuals

To run Median Polish, use the medpolish function on R

```
mp = medpolish(dat)
```

```
Overall: 148.7713
Row Effects:
       1        2        3        4        5        6        7        8        9       10       11
 -4.3975 416.8688 678.5188 -125.7012  51.8725 -59.1175 1050.9000 669.5225 -38.3225 -110.8350   0.0000
Column Effects:
       C1        C2        C3        T1        T2        T3
  3.65375  -2.19625 -31.96375  24.30625   2.19625 -47.67375
Residuals:
         C1       C2       C3      T1       T2        T3
1  -10.9475  23.7425    0.000   0.000 -63.5200    6.3200
2   33.7863  41.9762 -112.766  91.234 -33.7863  -38.9463
3   20.1363 156.3262  -70.416 137.584 -20.1363  -62.5963
4   -6.9638  -4.9638   31.184 -16.156   4.9638   38.1737
5   32.7825  -5.0275  -32.270   2.230  52.2100  -38.9500
6  -15.7275  40.7625   25.520 -26.780   0.0000    0.0000
7    9.2550  -9.2550 -348.797  55.203  77.1825 -192.6775
8  -62.5675 -17.6775  -15.420  14.880 354.5600   97.9000
9  -29.7225   2.1675   31.225 -12.575  -0.2950    0.2950
10   0.0000   0.0000   34.738 -19.063  -6.0125   32.0575
11   6.1550  -6.1550   19.103 -10.598  21.0825   -8.0775
```

To Calculate the samples gene expressions, add the overall median to the column effects

```
sampleExpresRob = mp$overall+ mp$col
```

Median Polish

We compare now the samples expressions obtain from LS
and Median Polish. The results show a big difference.

```
 sampleExpresLS
sampleC1 sampleC2 sampleC3 sampleT1 sampleT2 sampleT3
380.1973 396.6845 306.9355 422.6473 416.0182 315.8982

 sampleExpresRob
      C1       C2       C3       T1       T2       T3
152.4250 146.5750 116.8075 173.0775 150.9675 101.0975
```

We calculate the sum of squares differences standardized and in Percent.

```
SSdiff =  sum((sampleExpresRob-geneEffects)^2)/sum(sampleExpresRob^2)*100
[1] 271.377
```

This is clear evidence that the gene has many outliers.
If the gene expressions don't have outliers, we expect Ssdiff < 25%

Median Polish

Homework:

The dataset probdata.csv contains probe expressions of 1100 probes for 6 samples. The probes 1100 correspond to 100 different genes, 11 probes per gene. For each of the 100 genes:

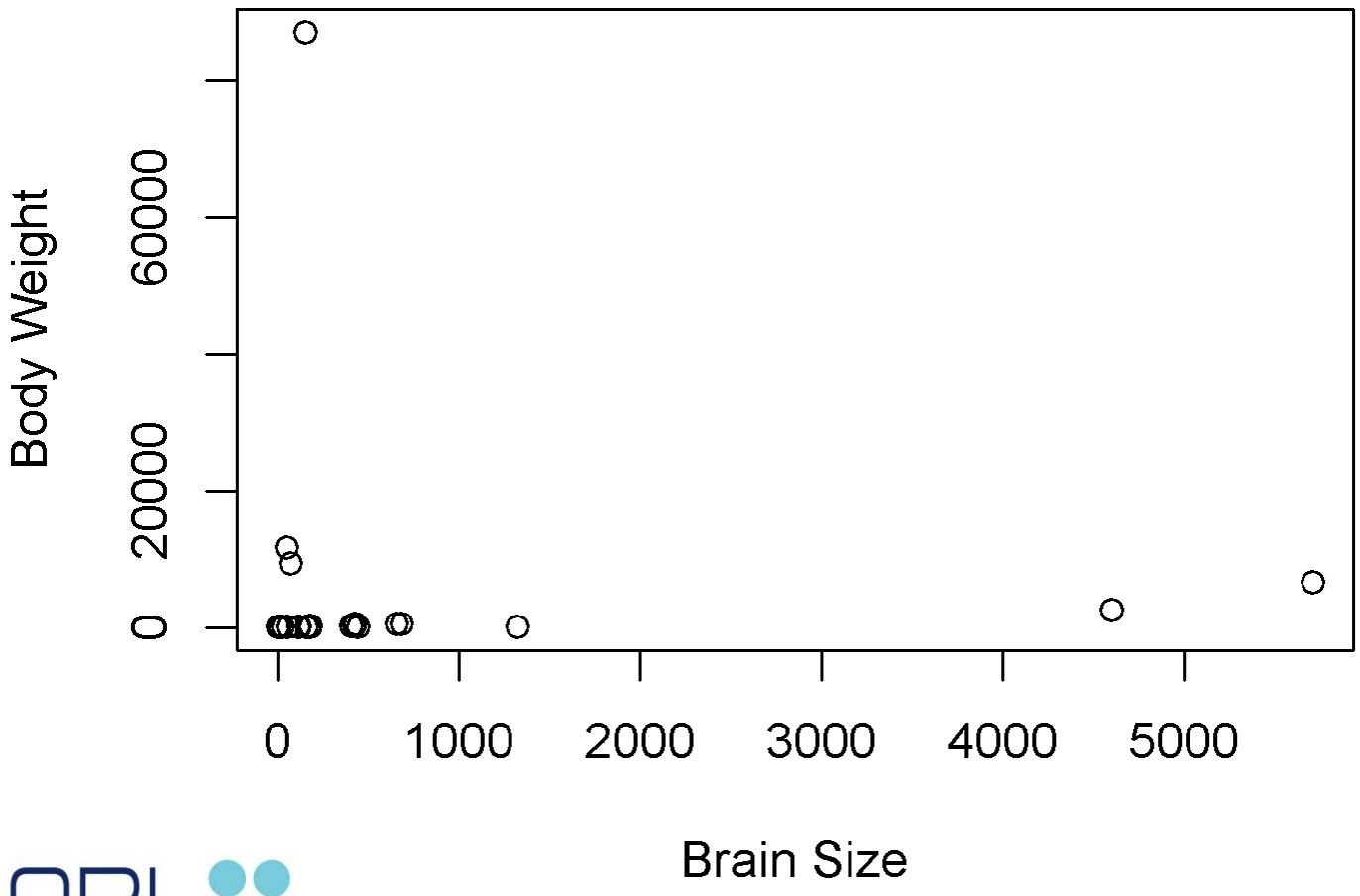
1. Calculate the LS expressions using the lm function as above

2. Calculate the Robust expressions using the medpolish function as above

3. Calculate the SSdiff statistic as above and calculate the proportion of genes that don't have outliers in the sense that Ssdiff is less than 25%

Class exercise (if there is time) or a homework

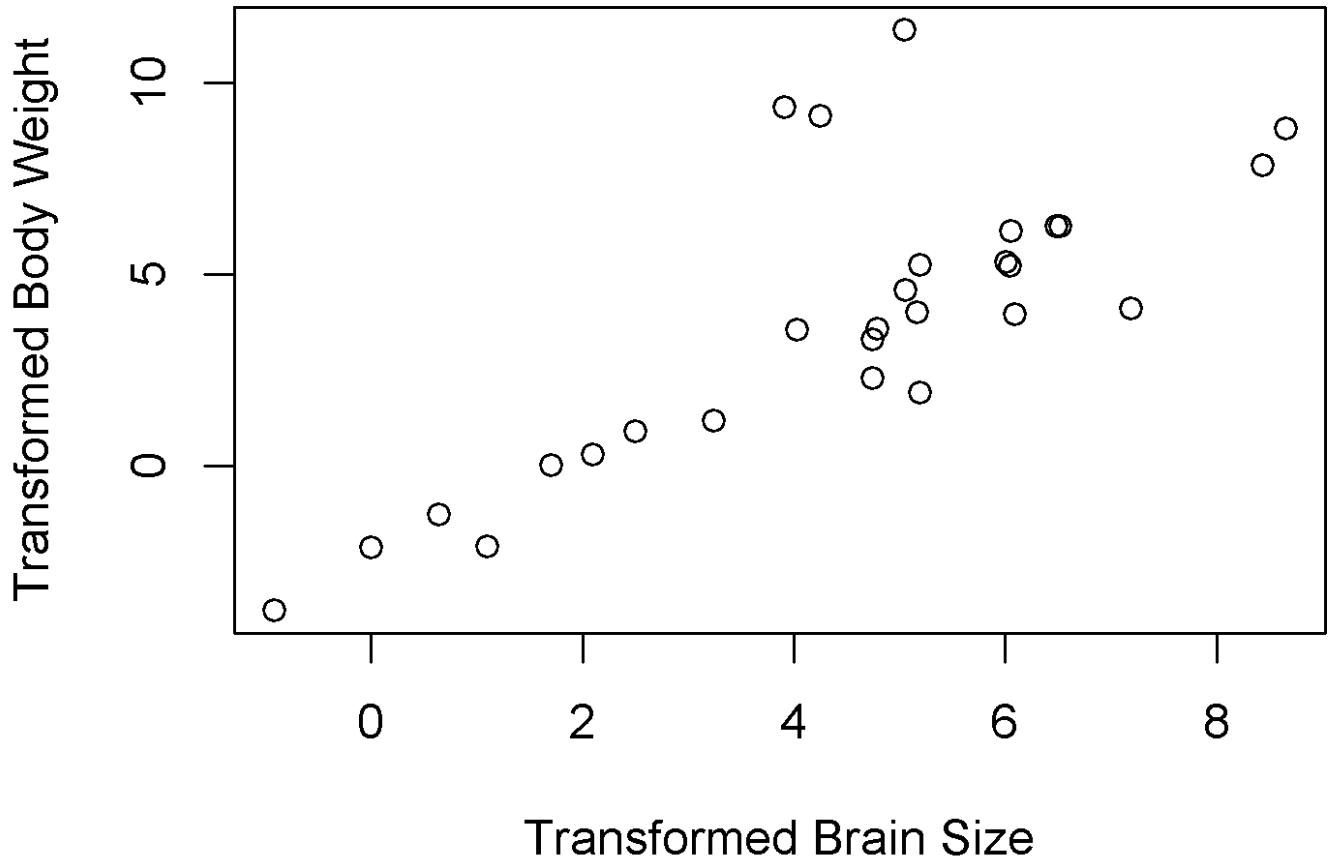
- Brain size vs. body mass (MASS::Animals)



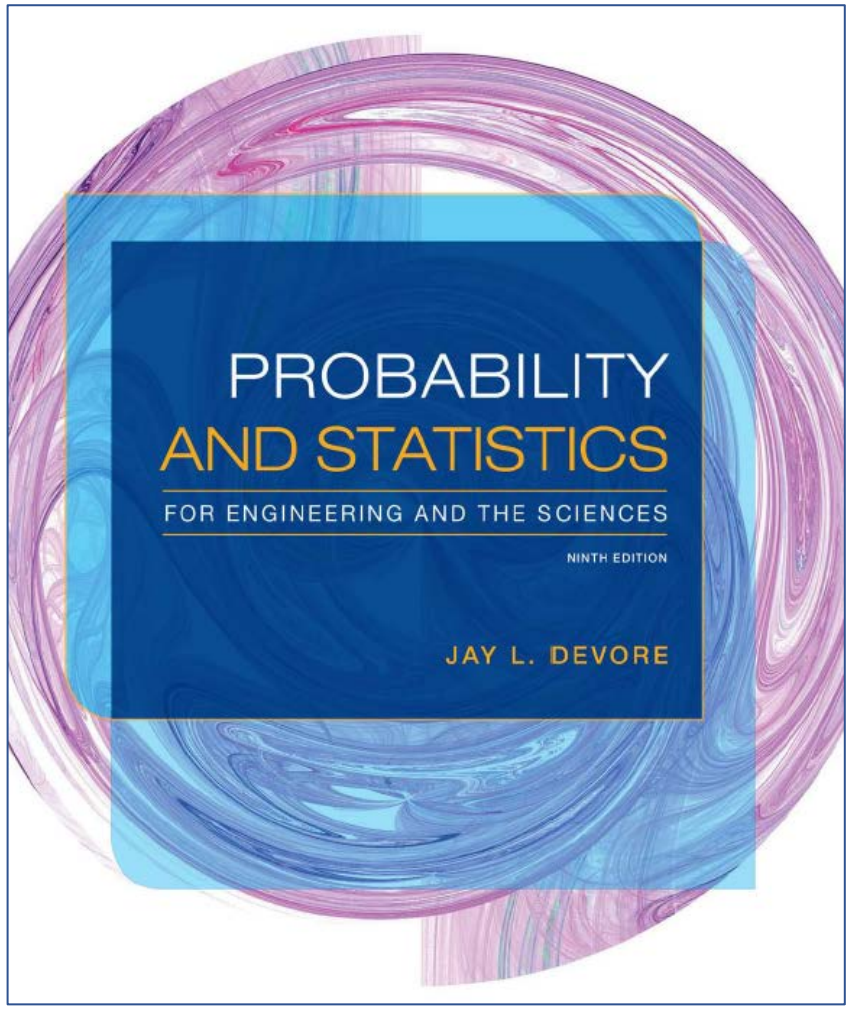
	Body Weight (kg)	Brain (g)
Mountain beaver	1.35	8.1
Cow	465	423
Grey wolf	36.33	119.5
Goat	27.66	115
Guinea pig	1.04	5.5
Dipliodocus	11700	50
Asian elephant	2547	4603
Donkey	187.1	419
Horse	521	655
Potar monkey	10	115
Cat	3.3	25.6
Giraffe	529	680
Gorilla	207	406
Human	62	1320
African elephant	6654	5712
Triceratops	9400	70
Rhesus monkey	6.8	179
Kangaroo	35	56
Golden hamster	0.12	1
Mouse	0.023	0.4
Rabbit	2.5	12.1
Sheep	55.5	175
Jaguar	100	157
Chimpanzee	52.16	440
Rat	0.28	1.9
Brachiosaurus	87000	154.5
Mole	0.122	3
Pig	192	180

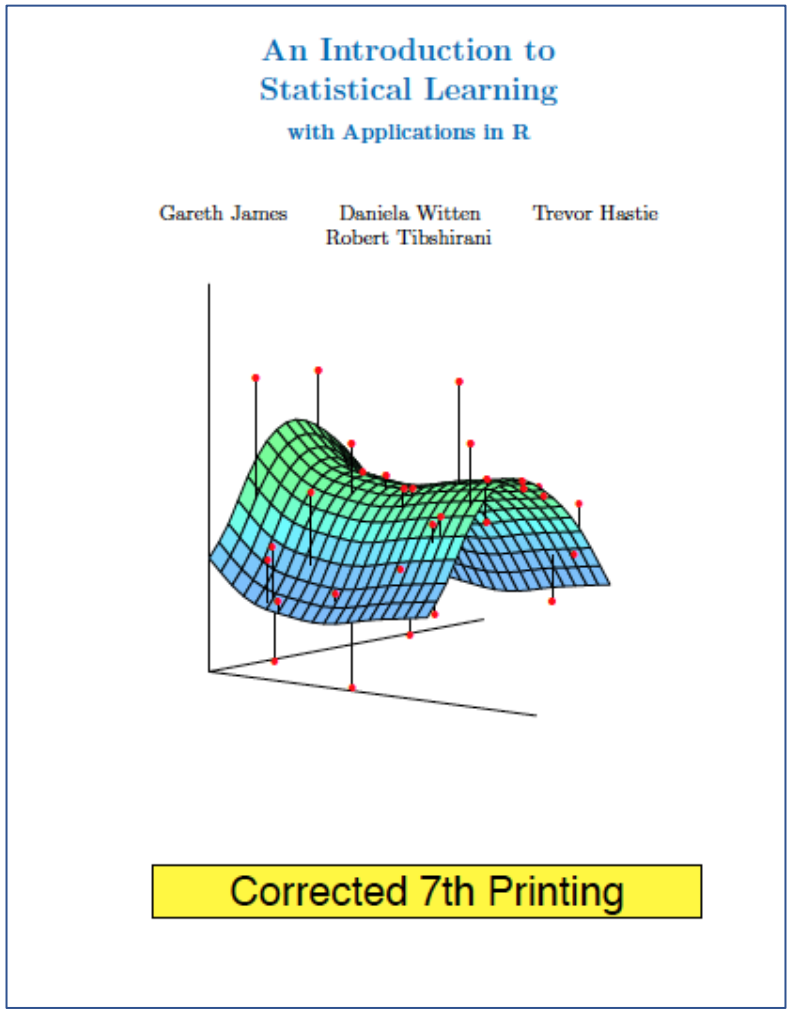
Class exercise (if there is time) or a homework

Use: transformations, outliers (justify!), lm, correlation, …



References





https://www.statlearning.com/