# Simple Linear Regression - Anscombe x1/y1 Exercise

SDS 291

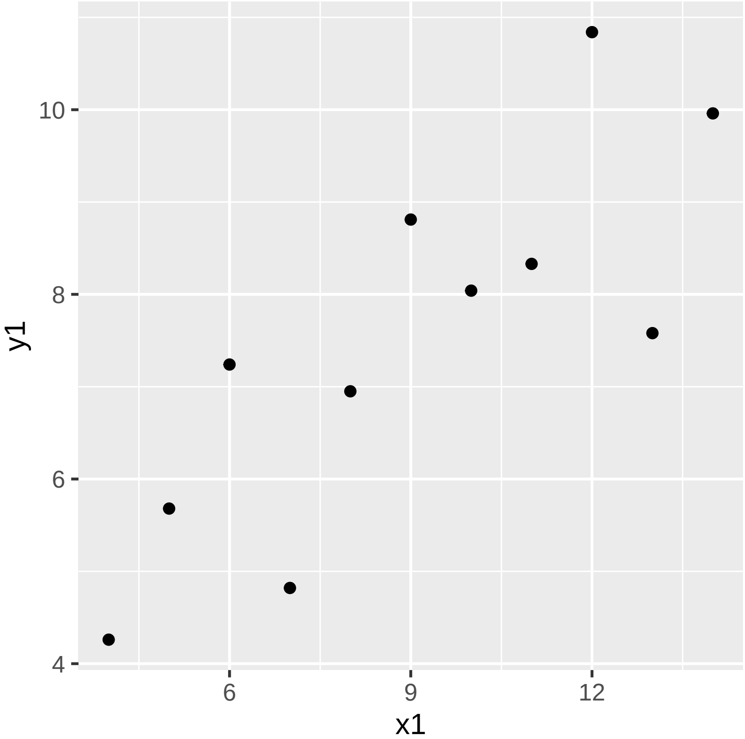
Monday, Feb 3, 2020

```
require(knitr)
require(tidyverse)
data("anscombe")
```

## Choose the Model

### Exercise 1

```
ggplot(data = anscombe, aes(x = x1, y = y1))  +
    geom_point()
```



This scatterplot demonstrates a positive, linear relationship between x1 and y1. There are no unusual observations, although it is difficult to tell with so few data points. The relationship is moderately strong, since you can see the points on the slope of the line, with a moderately steep slope

### Exercise 2

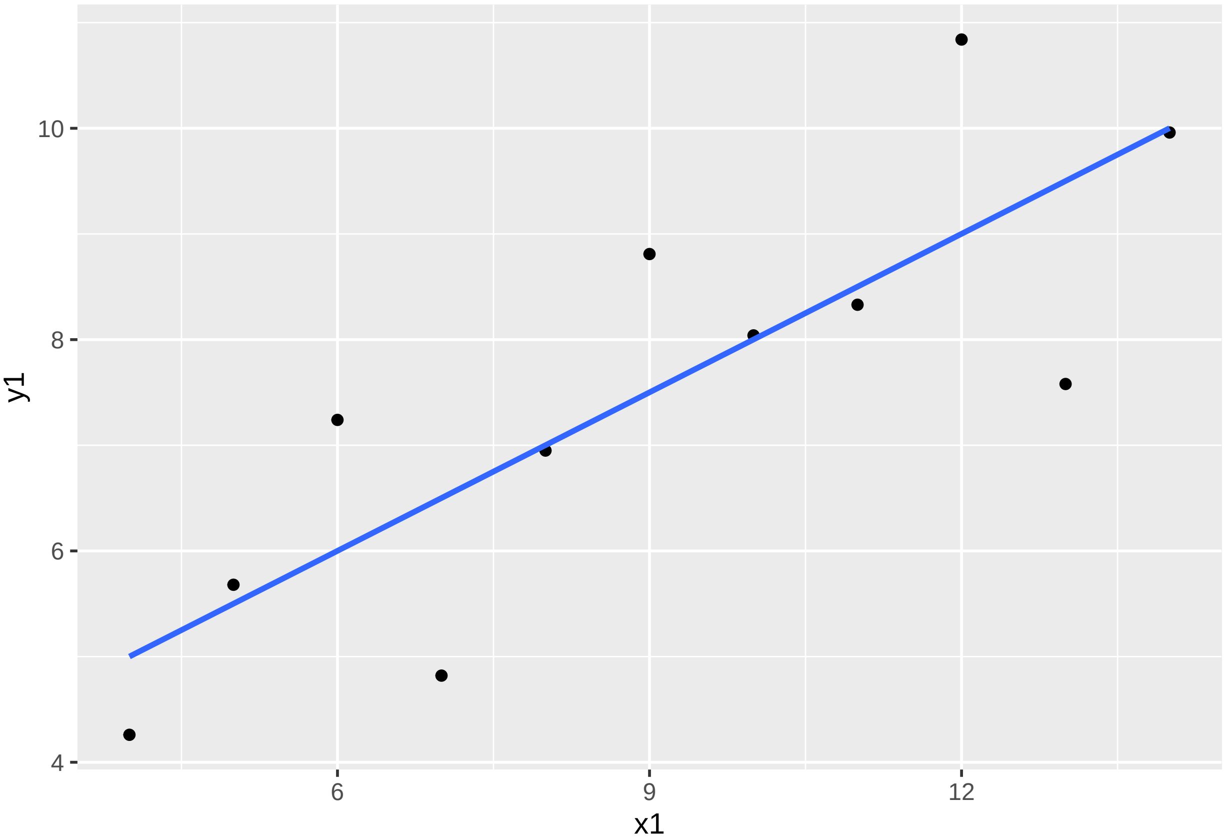
$$y = \beta_0 + \beta_1 X_1 + \epsilon$$

# Fit the model

```r
#fit the regression model
m1<-lm(y1 ~ x1 , data = anscombe)
summary(m1)
```

```
## 
## Call:
## lm(formula = y1 ~ x1, data = anscombe)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.92127 -0.45577 -0.04136  0.70941  1.83882 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)   
## (Intercept)   3.0001     1.1247   2.667  0.02573 * 
## x1            0.5001     0.1179   4.241  0.00217 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6665, Adjusted R-squared:  0.6295 
## F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

```r
#visualize the regression model
ggplot(data = anscombe, aes(x = x1, y = y1))  +
    geom_point() +            # Plot the scatterplot
    geom_smooth(method=lm,    # Add linear regression line
                se=FALSE)     # Don't add shaded confidence region
```



## Exercise 3

The fitted regression equation is:

$$y = \hat{\beta}_0 + \hat{\beta}_1 X_1$$

or

$$y = 3 + 0.5x1$$

## Exercise 4

The Intercept reflects the value of y1 when x1 is equal to 0. In other words, this model estimates that when x1 is 0, y1 will be equal to 3.

For each 1 unit* increase in x1, there is a 0.5 unit increase in y1.

*: Note that there are no real units for these data – but if there were, a complete answer would use those units and construct a sentence that really makes sense in english.

### Exercise 5

```
library(mosaic)
ans_predict<-makeFun(m1)
ans_predict(x1=6.5)
```

```
##        1 
## 6.250682
```

If x1 were 6.5, this model would esimate that y1 would be 6.25.

This checks out with the math: 3 + (0.5 x 6.5) = 3 + 3.25 = 6.25.

## Assess - Is a linear regression model a good fit for these data?

### Residuals and Predictions

```
library(broom)
m1_data<-augment(m1)
```

### Exercise 6

```
m1_data %>%
  filter(x1==8) %>%
  select(x1,.resid)
```

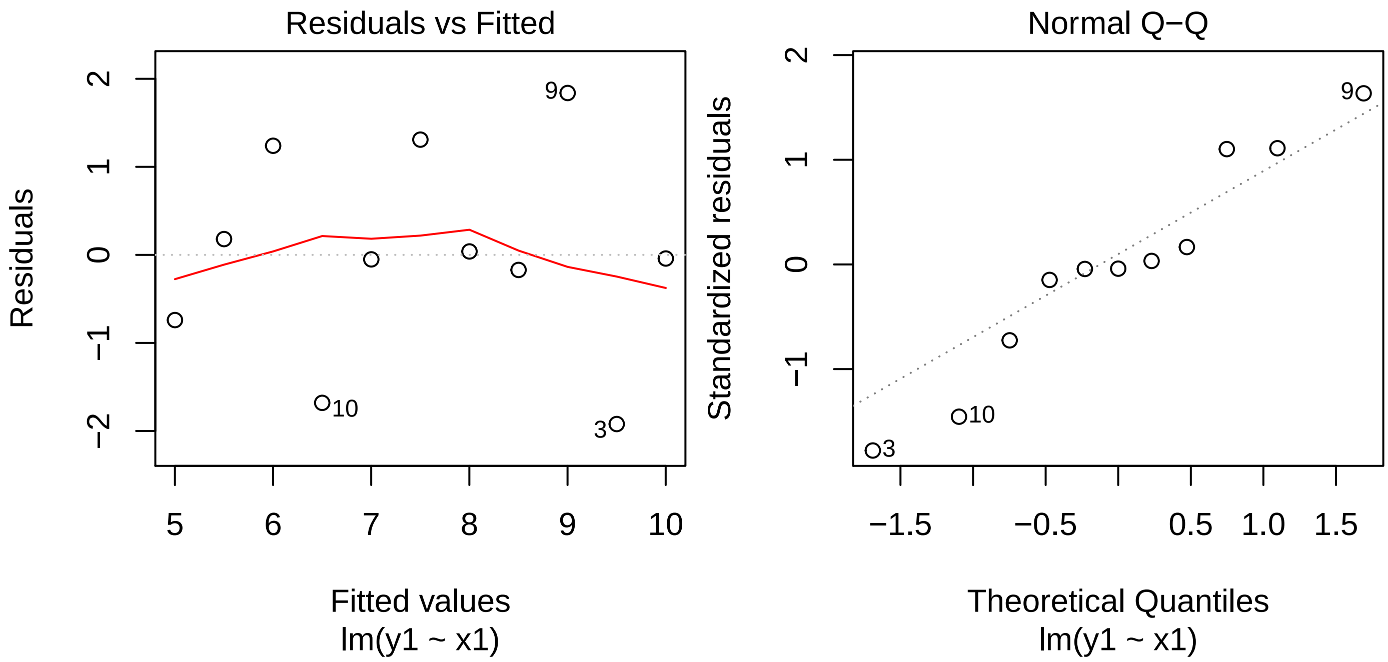
```
## # A tibble: 1 x 2
##      x1  .resid
##   <dbl>   <dbl>
## 1     8 -0.0508
```

If x1 = 8, the residual would be -0.05. In other words, the model would have over-estimated the expected value of y1 as 0.05 units higher than its actual value.

### Exercise 7

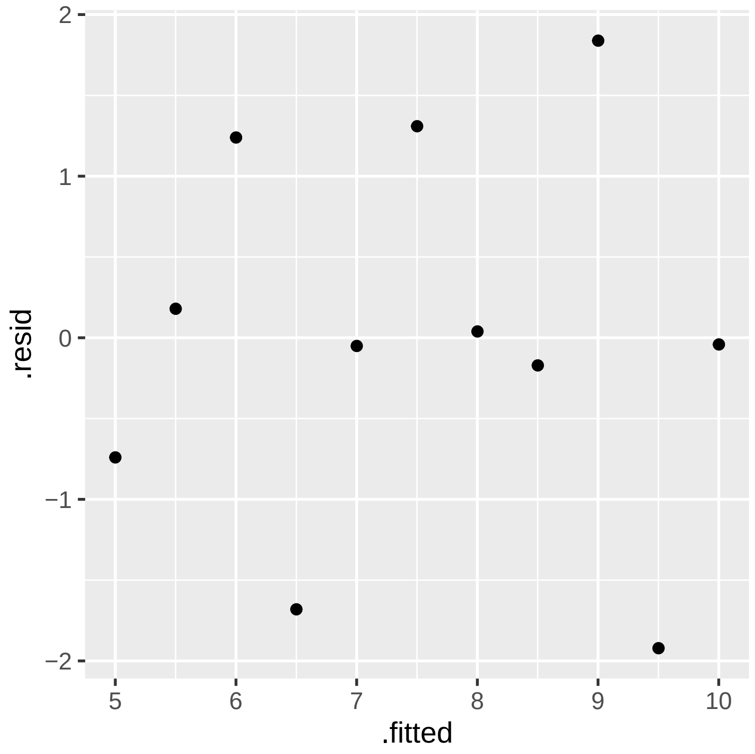
#### Base R

```r
plot(m1, which=c(1:2))
```

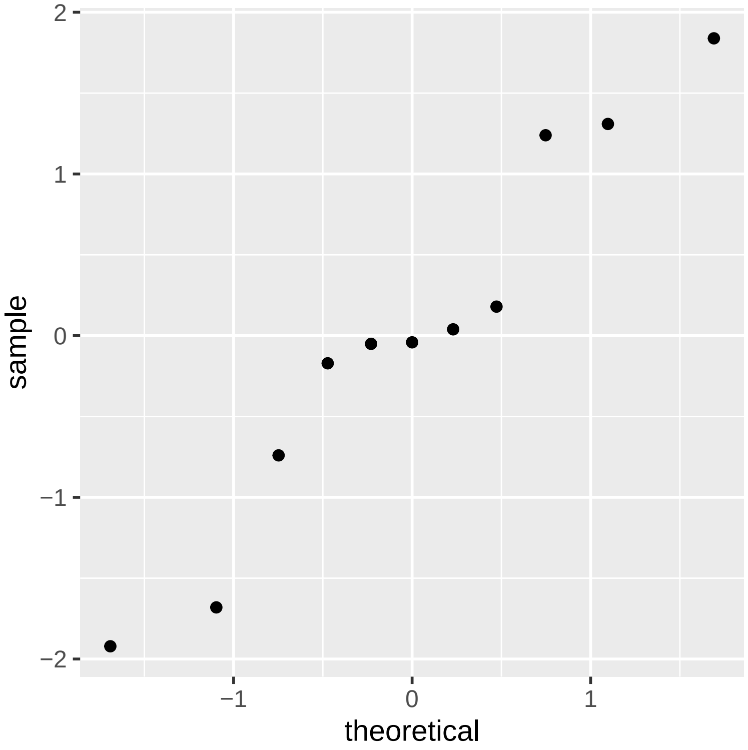


GGPlot

```r
ggplot(data = m1_data, aes(x = .fitted, y = .resid)) + geom_point()
```

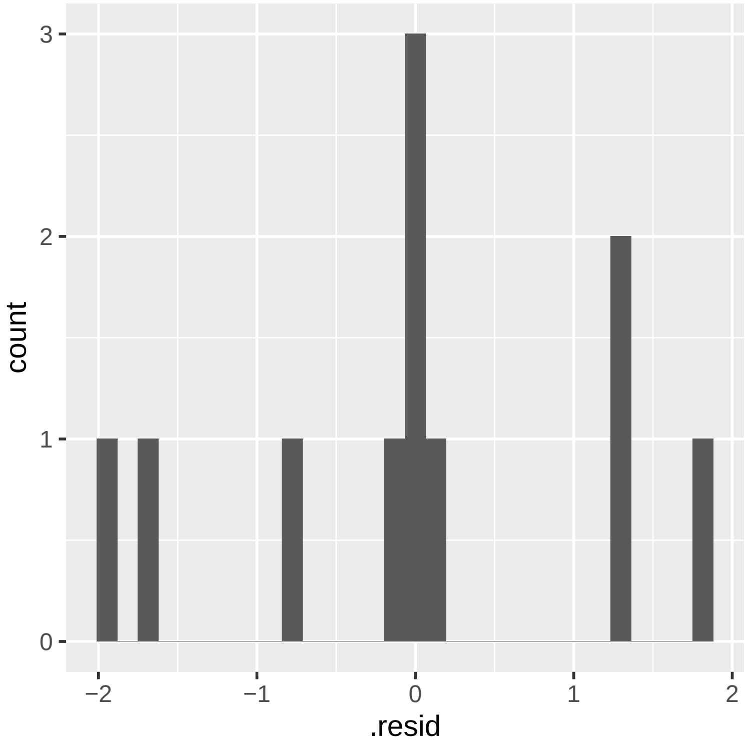


```r
ggplot(data = m1_data, aes(sample = .resid)) + geom_qq()
```



```r
ggplot(data = m1_data, aes(x = .resid)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



## Exercise 8

The regression assumptions seem to be reasonably met.

Linearity seems met, as evidenced by the lack of a persistant pattern in the data in the residual-fitted plot (i.e, there's no curve or linear trend left in the data). The red line in the "base" `R` plot is mostly flat and horizontal, suggesting that there is no shape to the residuals or that the linear model seems to fit these data well (if a linear model wasn't a good fit, there would be a distinct shape to this red line).

Constant / Equal Variance is met, since the residuals are fairly evenly distributed between -2 and 2 standardized residuals (the y-axis on the fitted-residual plot), which suggests the residuals are fairly equally distributed across the range of the fitted values.

Normality seems to be met from the bell-shaped, symmetric distribution of the histogram and that most of the points fall along the line of the Q-Q plot*, as they should if the residuals were distributed exactly along a normal distribution.

Zero Mean also appears to be met.

Indepdence and Randomness are unknown, since we don't know enough about the data.

* For more background on the QQPlot, see OpenIntro textbook pages 93-99 here)

## Exercise 9

$H_0 : \beta_1 = 0$

$H_A : \beta_1 \neq 0$

Because the p-value* for the slope (p = 0.00217) is below our significance level ($\alpha = 0.05$), we can reject the null hypothesis and conclude that there is a statistically significant, linear relationship between x1 and y1.

The p-value indicates the probability this relationship between x1 and y1 could be found due to chance alone. Since we have a relatively small p-value, corresponding to a small probability that the we would observe a relationship between x1 and y1 this size in a hypothetical world where there was no relationship between x1 and y1.

You can see the relationship between the test statistic and p-value with this applet for the t-distribution.

And see pages 68-77 in Open Intro (especially section 2.3.2 for interpretation of p-values) here

### Exercise 10

```r
confint(m1)
```

```
##                  2.5 %    97.5 %
## (Intercept) 0.4557369 5.5444449
## x1          0.2333701 0.7668117
```

We are 95% confident that the true relationship between x1 and y1 is between 0.23 and 0.77 in the population of interest*.

* Note that the population from which this sample is drawn is not clear. If this were a sample of Smith students, it would be that the relationship between x1 and y1 was between 0.23 and 0.77 among the total population of all Smith students.

* For more background on the confidence interval interpretaion, see OpenIntro textbook pages 102-107 here)

## More practice

### Exercise 11

The regression / residual standard error is calculated as the sum of squared errors (SSE) – literally squaring the residuals, then summing them – and, like other standard errors, taking square root of the value (SSE) that is divided by the degrees of freedom*

* The degrees of freedom are n-2: you have 11 observations, you used 2 of those to estimate the intercept and the slope, and you have 9 remaining degrees of freedom.

# Regression/residual standard error

$\hat{\sigma}_\epsilon = \sqrt{\frac{\sum(y-\bar{y})^2}{n-2}} = \sqrt{\frac{SSE}{n-2}}$

```r
SSE <- sum((m1_data$.resid)^2)
SSE
```

```
## [1] 13.76269
```

```r
RSE<-sqrt(SSE/9)
RSE
```

```
## [1] 1.236603
```

1. Do you get the same answers as what's reported in the linear model results and ANOVA table (we'll talke more abotu the ANOVA table next week – just see if you can find the same number for now)

```r
summary(m1)
```

```
##
## Call:
## lm(formula = y1 ~ x1, data = anscombe)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92127 -0.45577 -0.04136  0.70941  1.83882
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0001     1.1247   2.667  0.02573 *
## x1            0.5001     0.1179   4.241  0.00217 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6665, Adjusted R-squared:  0.6295
## F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

```r
anova(m1)
```

```
## Analysis of Variance Table
##
## Response: y1
##           Df Sum Sq Mean Sq F value  Pr(>F)
## x1         1 27.510 27.5100   17.99 0.00217 **
## Residuals  9 13.763  1.5292
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We got the same residual standard error as the regression summary output: 1.237 on 9 degrees of freedom.

We also got the same value of the sum of squared residuals as is included in the ANOVA table: 13.763.

### Exercise 13

The regression standard error can be interpreted as the "average" error in the model, and is in the units of the y / response variable. In this case, the model is off by, on average, 1.3 units of y1. When the units are clearer, this interpretation makes more sense.

To practice,

1. check what the regression/residual standard error was from the Cereal problem from homework and interpret in a sentence, and
2. read the part of chapter 1 that includes the interpretation of the regression standard error.