Topic 3
Correlation and Regression
Linear Regression II

Outline

Example - *de novo* Mutations

Example. We continue to investigate the relationship of age of parents to the *de novo* mutations in the offspring for the 78 Icelandic trios. We use the age of the parents to predict the number of mutations in the offspring. Thus, age is on the horizontal axis.

Example - *de novo* Mutations

We can quickly obtain the regression line using R.

```
> lm(mutations~age)

Call:
lm(formula = mutations ~ age)

Coefficients:
(Intercept)          age
      2.815        2.125
```

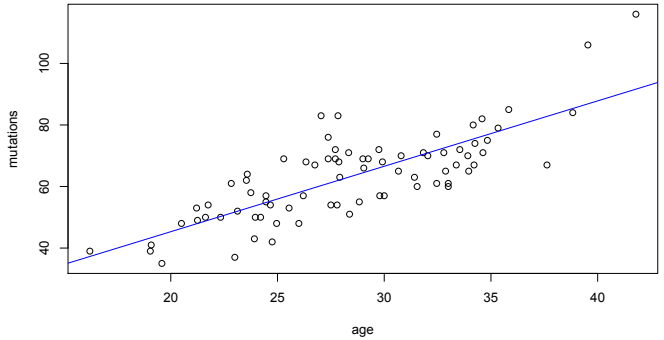
Thus, the regression line has the equation.

$$\widehat{\text{mutations}} = 2.815 + 2.125\ \text{age}.$$

Example - *de novo* Mutations

For more advanced analysis, we store the results of regression. Here we plot the data and add the regression line to the plot.

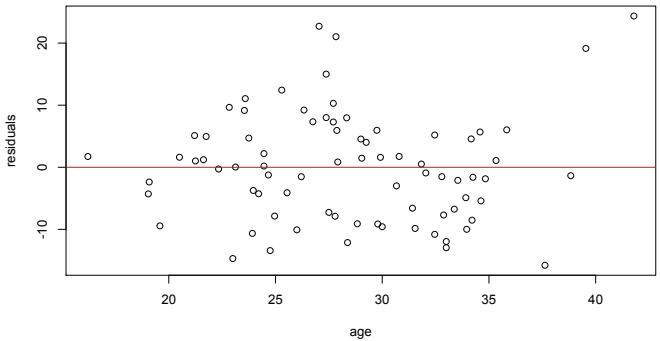
```
> mutations.lm<-lm(mutations~age)
> plot(age,mutation)
> abline(mutations.lm,col="blue")
```



Example - *de novo* Mutations

Next, we ask for the residuals, make a residual plot, and create a horizontal line at 0.

```
> residuals<-resid(mutations.lm)
> plot(age,residuals)
> abline(h=0,col="red")
```



Example - *de novo* Mutations

We use the regression line to predict the number of mutations for parents whose average age is 20, 30, or 40.

```
> agepredict<-c(20,30,40)
> predictions<-predict(mutations.lm,newdata=data.frame(age=agepredict))
> data.frame(agepredict,predictions)
  agepredict predictions
1         20    45.32367
2         30    66.57824
3         40    87.83281
```

Exercise. Verify the predictions above by hand.

Example - *de novo* Mutations

For a general summary,

```
> summary(mutations.lm)
Call:
lm(formula = mutations ~ age)
Residuals:
     Min       1Q   Median       3Q      Max
-15.7849  -7.1364  -0.1244   5.1745  24.3591
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.8145     5.5034   0.511    0.611
age           2.1255     0.1904  11.164   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 8.79 on 76 degrees of freedom
Multiple R-squared:  0.6212, Adjusted R-squared:  0.6162
F-statistic: 124.6 on 1 and 76 DF,  p-value: < 2.2e-16
```

R-squared

For explanatory variable x and response variable y, recall the definition of correlation

$$r = \frac{\text{cov}(x, y)}{s_x s_y}, \quad \text{cov}(x, y) = r s_x s_y.$$

We can use this to give and expression for the slope of the regression line

$$\hat{\beta} = \frac{\text{cov}(x, y)}{\text{var}(x)} = \frac{r s_x s_y}{s_x^2} = r \frac{s_y}{s_x}.$$

Then, write the regression line in point-slope form

$$\hat{y} - \bar{y} = r \frac{s_y}{s_x}(x - \bar{x}).$$

Now take the variance of both sides, using the quadratic identity

$$s^2_{\text{FIT}} = \text{var}(\hat{y}) = \text{var}\left(r \frac{s_y}{s_x}(x - \bar{x}) \right) = r^2 \frac{s_y^2}{s_x^2} s_x^2 = r^2 s_y^2 = r^2 s^2_{\text{DATA}}.$$

R-squared

$$s^2_{\text{FIT}} = r^2 s^2_{\text{DATA}}$$

For the mutation data,

```
> cor(age,mutations)^2
[1] 0.621215
```

We say that 62% of the variation in the number of *de novo* mutations can be explained by the average age of the parents.

The FIT and RESIDUALS are uncorrelated.

$$\begin{aligned} s^2_{DATA} &= s^2_{FIT} + s^2_{RESID} \\ &= r^2 s^2_{DATA} + (1 - r^2) s^2_{DATA}. \end{aligned}$$

In this case, we say that

- r^2 of the variation in the response variable is due to the fit and
- the rest $1 - r^2$ is due to the residuals.