Inference for regression

Load libraries and set theme

```
library(dplyr)
library(ggplot2)

theme_set(theme_bw(base_size = 18))
```

load twin iq data and have a look at the first few rows

```
twins <- read.csv("https://dyurovsky.github.io/psyc20100/data/demos/twins.csv")
head(twins)
```

```
##   twin_a twin_b
## 1     68     63
## 2     94     86
## 3     93     99
## 4    115    101
## 5    104    114
## 6     71     76
```

find the correlation between the iq of twin a and twin b

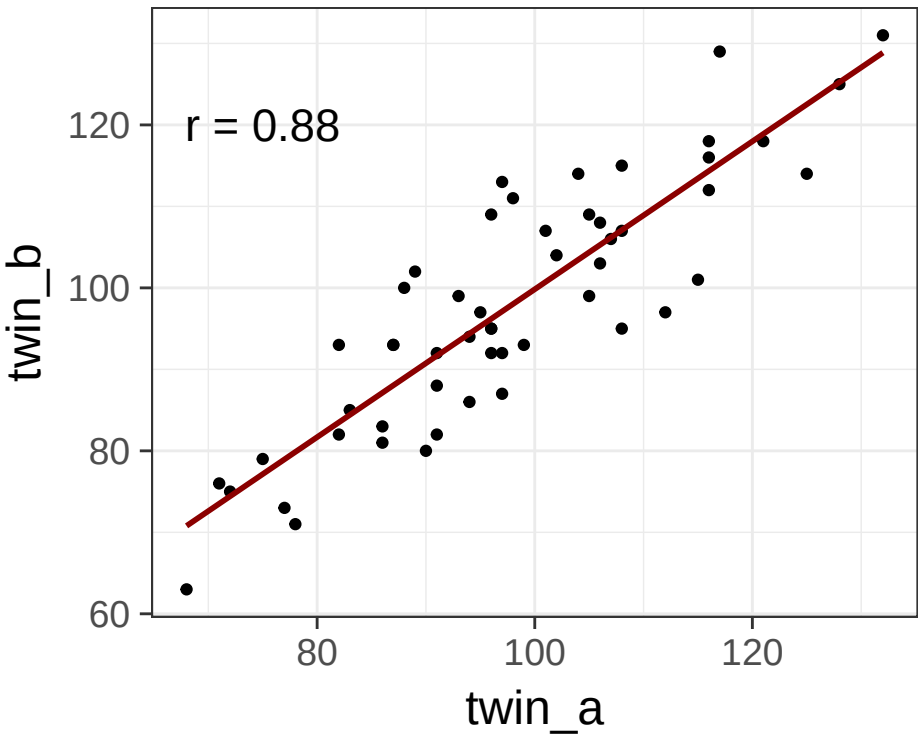
```
twin_cor <- cor(twins$twin_a, twins$twin_b)

twin_cor
```

```
## [1] 0.8757779
```

make a scatterplot of the relationship between twin a and b iqs. Also plot a regression line and annotate with the correlation coefficient.

```
qplot(data = twins, x = twin_a, y = twin_b) +
  geom_smooth(method = "lm", se = FALSE, color = "darkred") +
  annotate("text", x = 75, y = 120, size = 6,
           label = paste0("r = ", round(twin_cor, digits = 2)))
```



make a linear regression predicting twin b from twin a

```
twin_lm <- lm(twin_b ~ twin_a, data = twins)
summary(twin_lm)
```

```
##
## Call:
## lm(formula = twin_b ~ twin_a, data = twins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.717  -5.365  -0.180   4.635  15.894
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.08670    6.92036   1.313    0.195
## twin_a       0.90741    0.07004  12.957   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.417 on 51 degrees of freedom
## Multiple R-squared:  0.767,  Adjusted R-squared:  0.7624
## F-statistic: 167.9 on 1 and 51 DF,  p-value: < 2.2e-16
```

the slope of a regression line is equal to the correlation when the independent and dependent variabels are both standardiZed and no slope is fit

```
scale_twins <- twins %>%
  mutate_each(funs(scale), twin_a, twin_b)

scale_twin_lm <- lm(twin_b ~ twin_a + 0, data = scale_twins)
```

```r
summary(scale_twin_lm)
```

```
##
## Call:
## lm(formula = twin_b ~ twin_a + 0, data = scale_twins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90142 -0.35257 -0.01183  0.30458  1.04448
##
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## twin_a  0.87578    0.06694   13.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4827 on 52 degrees of freedom
## Multiple R-squared:  0.767,  Adjusted R-squared:  0.7625
## F-statistic: 171.2 on 1 and 52 DF,  p-value: < 2.2e-16
```

```r
twin_cor
```

```
## [1] 0.8757779
```

That's becasue the slope of a regression line is r * sd_x / sd_y

```r
twin_cor * sd(twins$twin_b)/sd(twins$twin_a)
```

```
## [1] 0.9074141
```

```r
coef(twin_lm)["twin_a"]
```

```
##    twin_a
## 0.9074141
```