Simple Linear Regression

Youmi Suk

School of Data Science, University of Virginia

2. 14. 2022

Overview

1. Motivating Example

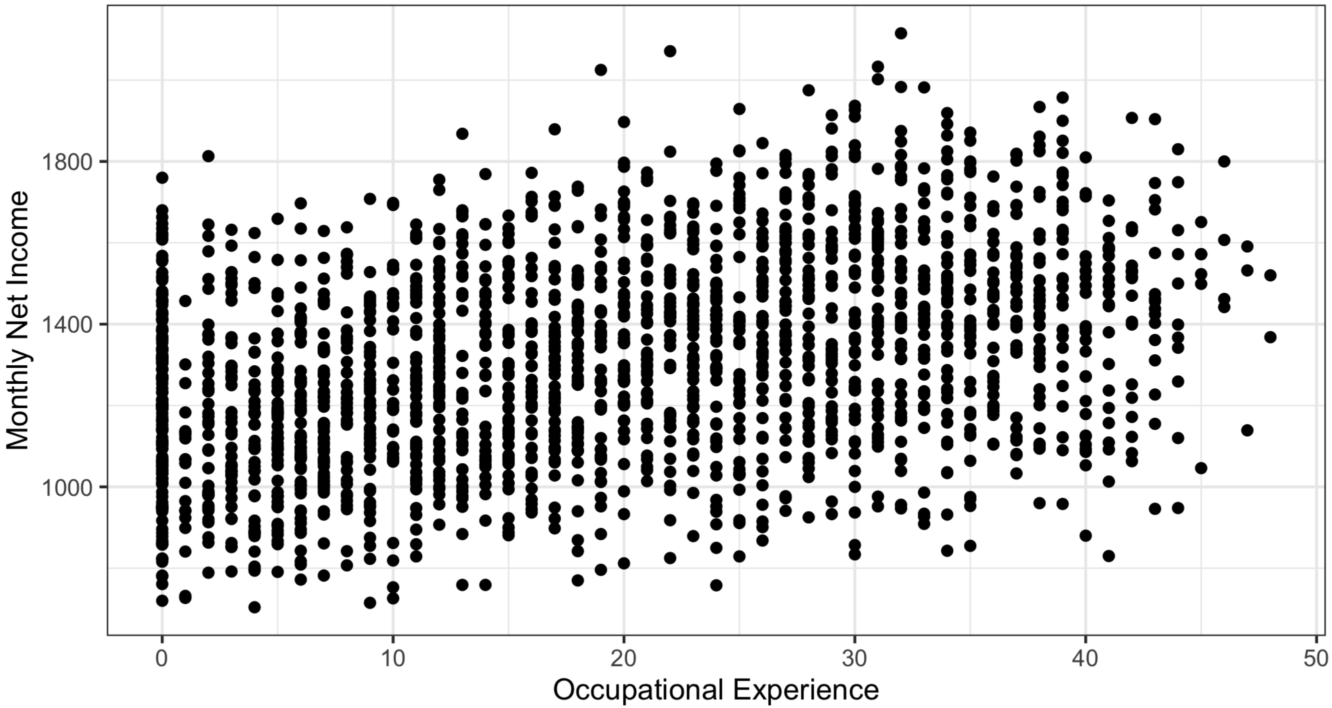
2. Simple Linear Regression

- Compute Regression Coefficients
- Compute Predicted Values

3. Scatterplots with Regression Lines and Predicted Values

Motivating Example

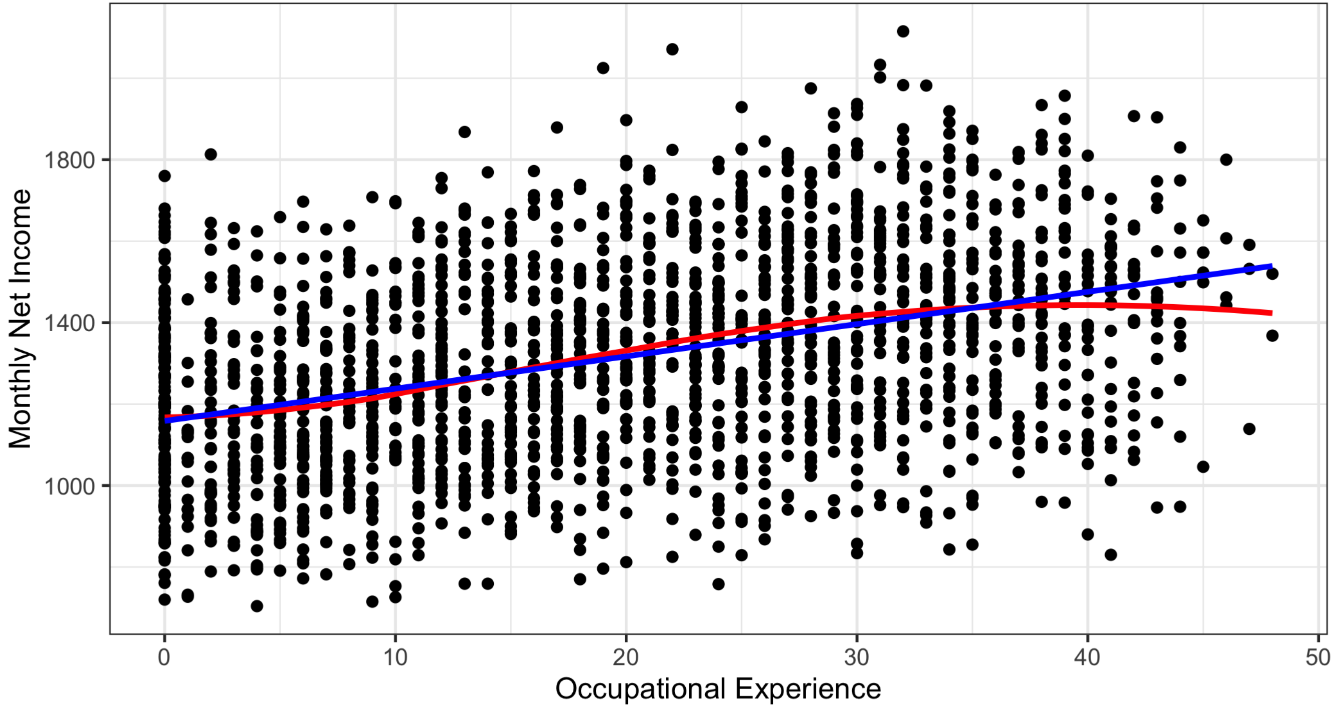
- Income and Occupational Experience:



- *Example*: simulated data of monthly net income (1922 observations)

Example: Income ~ Occup. Experience

- Scatterplot with `loess` line and `lm` line:



Simple Linear Regression

Though nonlinear relations between X and Y are quite common in practice, we start by investigating the simple linear relationship.

That is, we assume that the mean values of Y linearly increase or decrease with increasing values of X.

The linearity assumption allows us to formulate the relationship between the dependent variable (Y) and the independent variable (X) in an algebraic form.

The path of means is given by the linear equation: $\overline{Y}(X) = A + BX$

In regression analysis, we use Y 'hat' notation: $\hat{Y}(X) = A + BX$

Simple Linear Regression

However, almost all individual observations deviate from the linear path of means. This is due to an error term E_i (measurement error; influential covariates other than X). Hence, the dependent value (Y_i) for the i-th observation is given by

$$Y_i = A + BX_i + E_i$$

with predicted values (= conditional mean values = regression line)

$$\hat{Y}_i = A + BX_i$$

Y_i = dependent variable; X_i = independent variable;

$\hat{Y}_i$ = predicted/fitted value

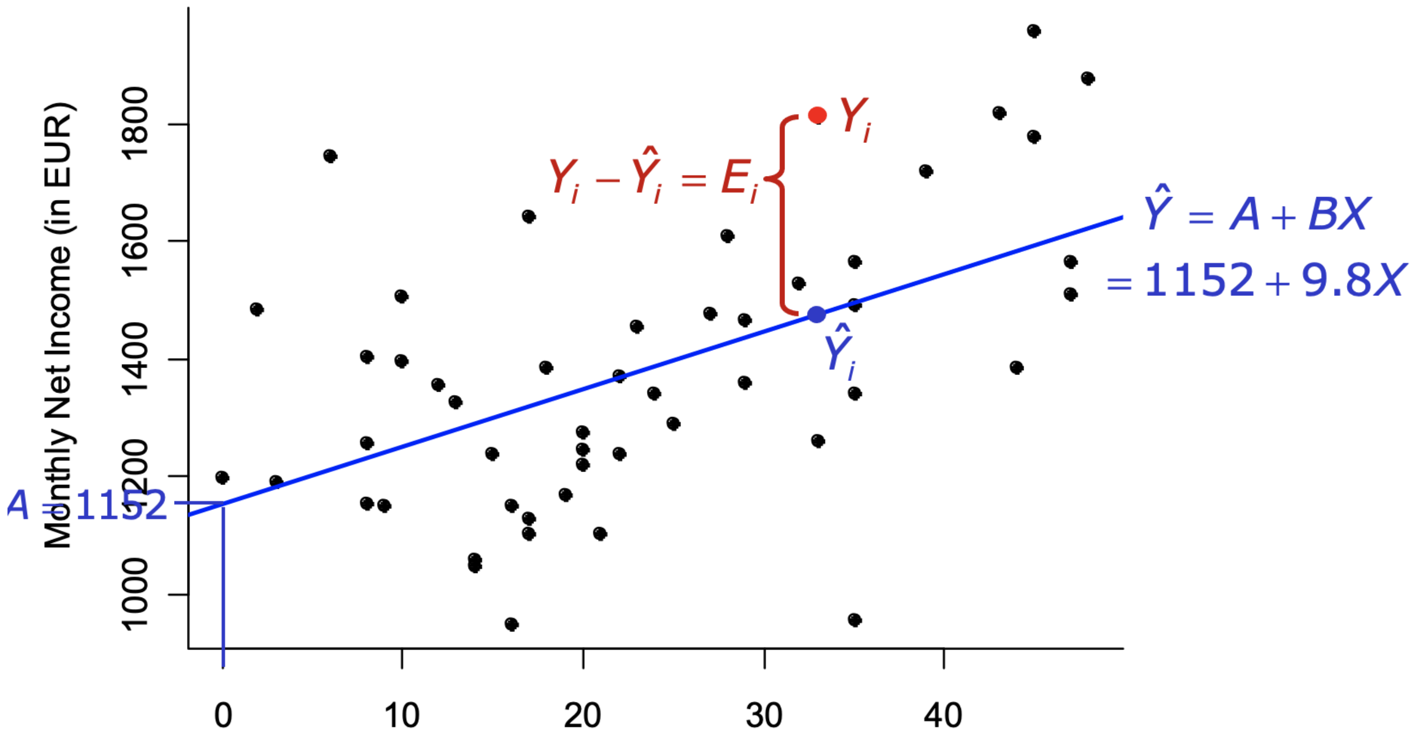
A = intercept; B = slope;

E_i = error term (residual)

Terminology

Y	**X**
Dependent Variable	Independent Variable
Explained Variable	Explanatory Variable
Reponse Variable	Control Variable
Predicted Variable	Predictor
Regressand	Regressor
Outcome	Covariate/Variable

Regression Line



Regression Equation: Meaning

A … intercept: mean value of Y for $X = 0$

$$\hat{Y}_i = A + BX_i = A + B \cdot 0 = A$$

B … slope: if X increases by one unit, Y increases (decreases) on overage by B units:

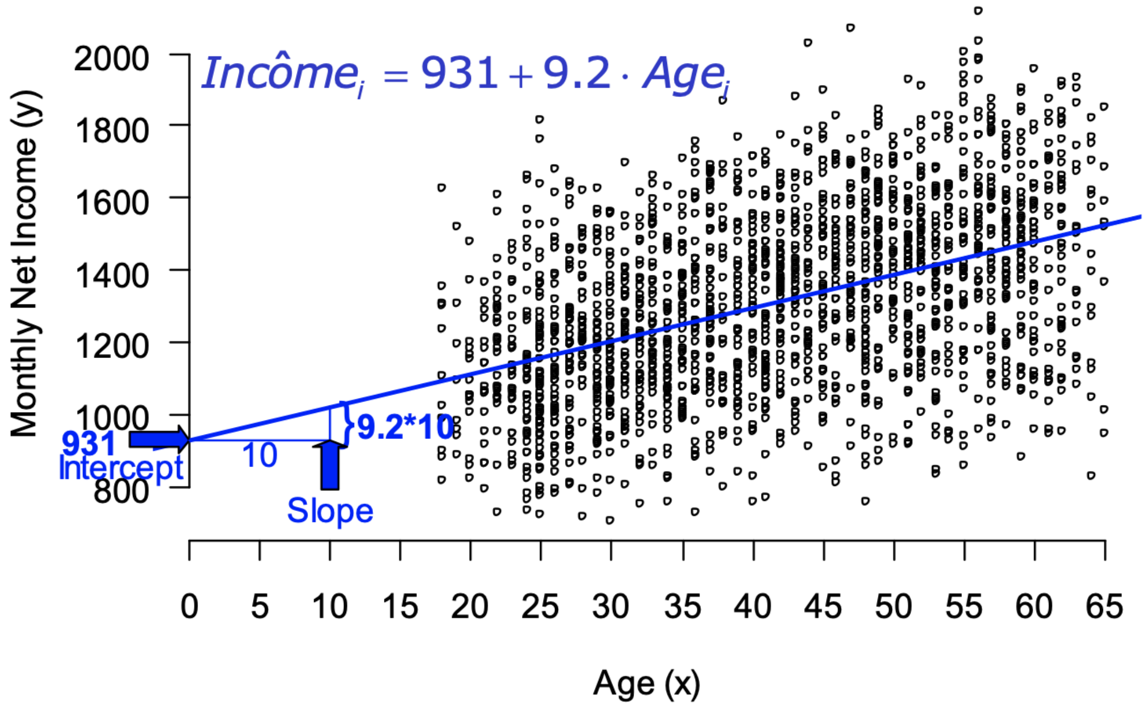
$$\hat{Y}|(X+1) - \hat{Y}|X = (A + B(X+1)) - (A + BX)$$

$$= A + BX + B - A - BX = B$$

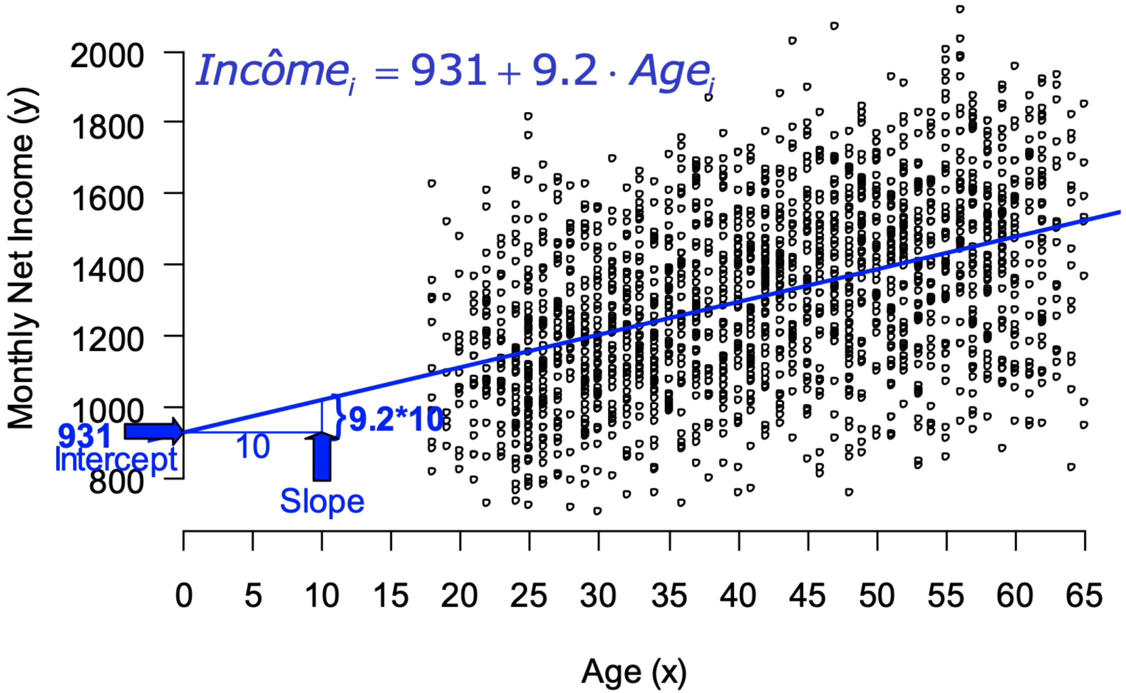
or by using the first derivative:

$$\frac{d\hat{Y}}{dX} = \frac{d(A+BX)}{dX} = B$$

Interpretation: *income ~ age*



Interpretation: *income ~ age*



A: The *mean* income of a person of age 0 is 931 EUR
(not really meaningful)

B: An increase in age by one year increases the *mean* income by 9.2 EUR.
An increase in age by 10 years increases the *mean* income by 92 EUR.

Estimating the Linear Regression Line

We need an optimization criterion in order to determine the regression coefficients of the regression equation: $Y_i = A + BX_i + E_i$ ('regression of Y on X'):

$$\text{Least Squares Criterion}$$

$$\text{(Ordinary Least Squares, OLS)}$$

Minimize the sum of squared residuals, i.e., minimize the sum of squared deviations of observed values from predicted values.

$$\sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - (A + BX_i))^2 \rightarrow \min_{A,B}$$

Estimated Regression Coefficients

Using the least squares principle, we get the following regression coefficients:

- slope: $B = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{S_{XY}}{S_X^2}$
- intercept: $A = \overline{Y} - B\overline{X}$

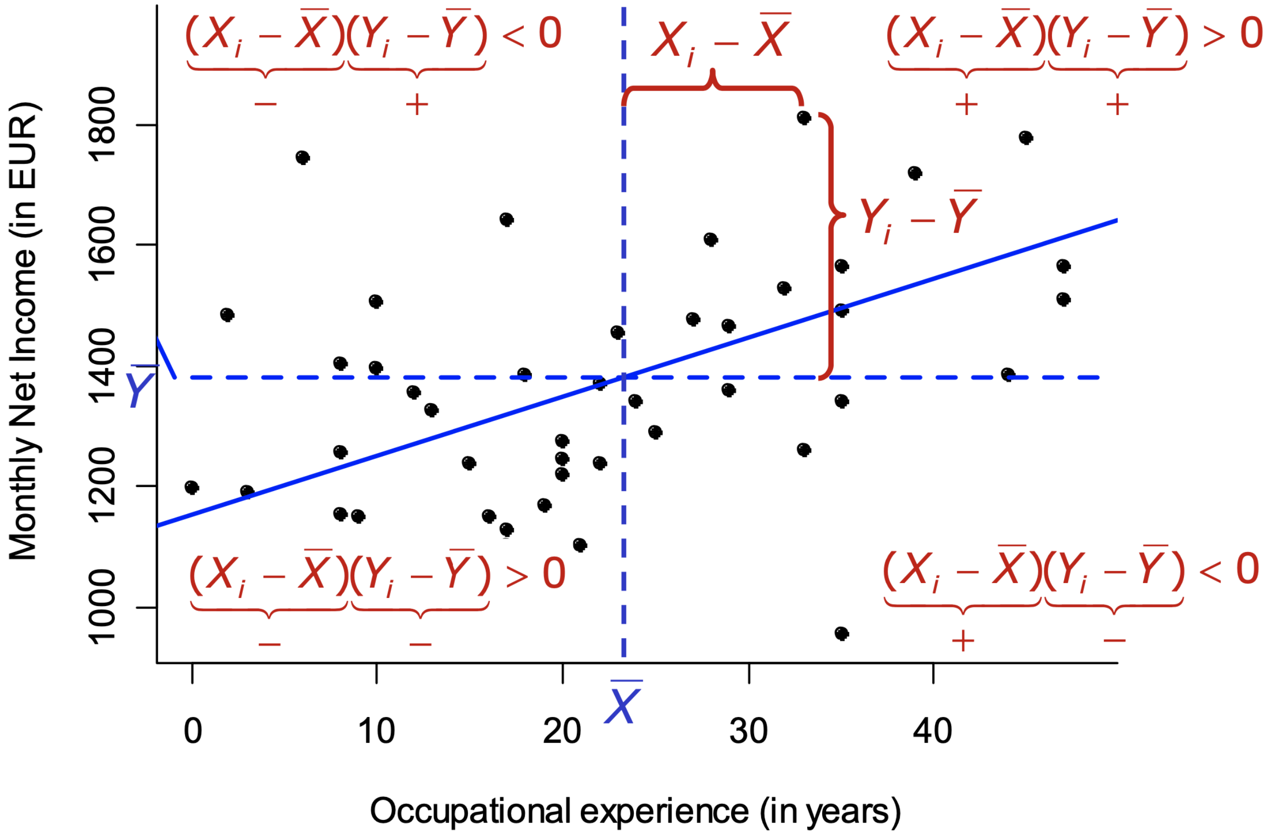
with

the sample means $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$;

the sample standard deviation $S_X = \sqrt{\frac{\sum(X_i - \bar{X})^2}{N-1}}$

and the sample covariance $S_{XY} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{N-1}$

Covariance



$\underbrace{(X_i - \bar{X})}_{-}\underbrace{(Y_i - \bar{Y})}_{+} < 0$
$X_i - \bar{X}$
$\underbrace{(X_i - \bar{X})}_{+}\underbrace{(Y_i - \bar{Y})}_{+} > 0$
$Y_i - \bar{Y}$
$\bar{Y}$
$\underbrace{(X_i - \bar{X})}_{-}\underbrace{(Y_i - \bar{Y})}_{-} > 0$
$\underbrace{(X_i - \bar{X})}_{+}\underbrace{(Y_i - \bar{Y})}_{-} < 0$
$\bar{X}$
Monthly Net Income (in EUR)
1000
1200
1400
1600
1800
0
10
20
30
40
Occupational experience (in years)

Covariance

- S_{XY} measures the co-variation of X and Y, i.e., to what extent and in which direction does Y co-vary with X?
    - $S_{XY} > 0$: positive relation (Y increases with increasing X); slope of the regression line is positive.
    - $S_{XY} < 0$: negative relation (Y decreases with increasing X); slope of the regression line is negative.
    - $S_{XY} = 0$: no relation (Y varies independent of X; X is independent of Y); slope of the regression line is zero.
- Note that $S_{XY} = S_{YX}$ and $S_{XX} = S_X^2$.

Correlation

- The covariance depends on the units of measurement and is frequently not easy to interpret.
- However, we can use the covariance to construct a standardized measure that indicates the strength of the linear relationship between two continuous variables X and Y: the correlation coefficient r_{XY}.

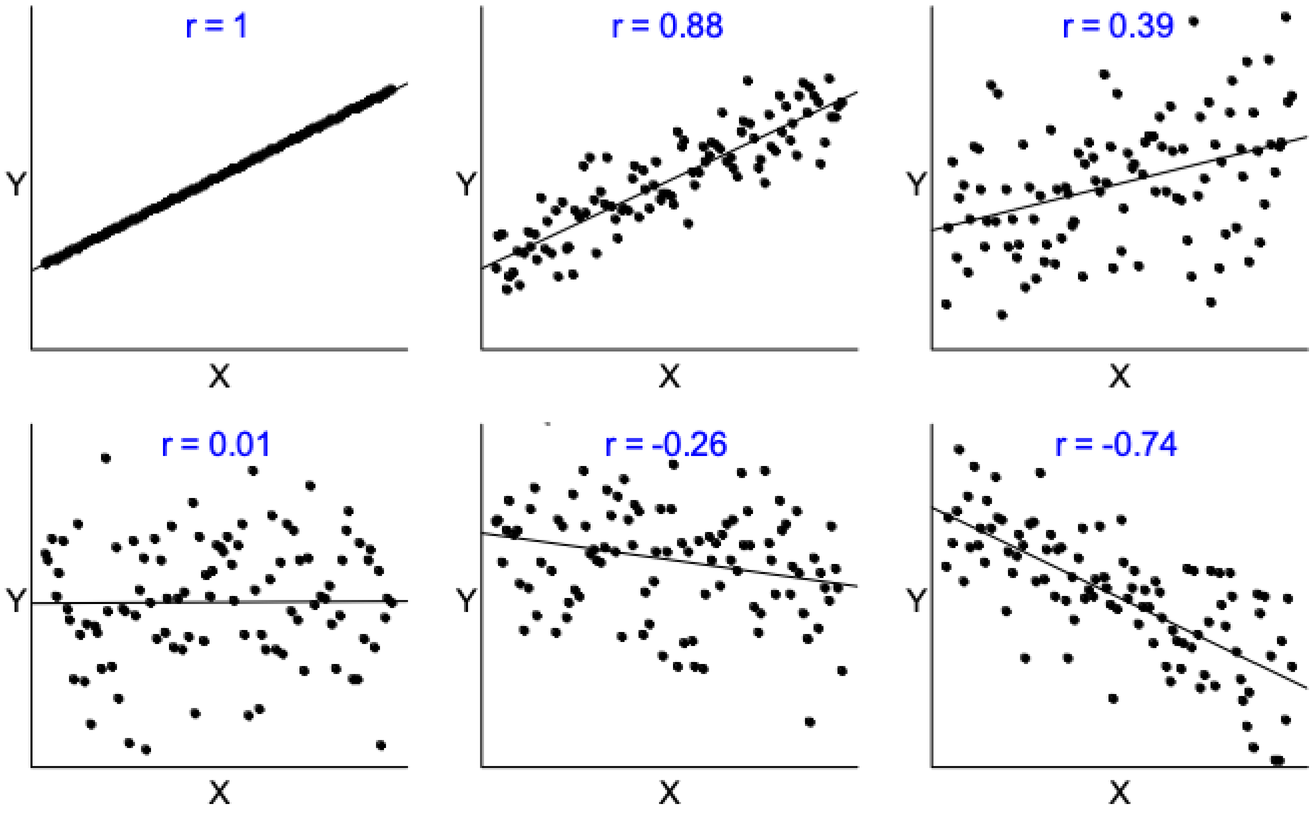
$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{N-1}}{\sqrt{\frac{\sum(X_i - \bar{X})^2}{N-1}}\sqrt{\frac{\sum(Y_i - \bar{Y})^2}{N-1}}}$$

$$= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}}$$

Correlation

- The correlation coefficient represents a standardized covariance and is always between –1 and +1.
- We use the correlation coefficient for assessing the strength of the linear relationship between X and Y:
  - $r_{XY} = 1$ indicates a perfect positive linear relationship.
  - $r_{XY} = -1$ indicates a perfect negative linear relationship.
  - $r_{XY} = 0$ indicates that there is no (linear) relationship.

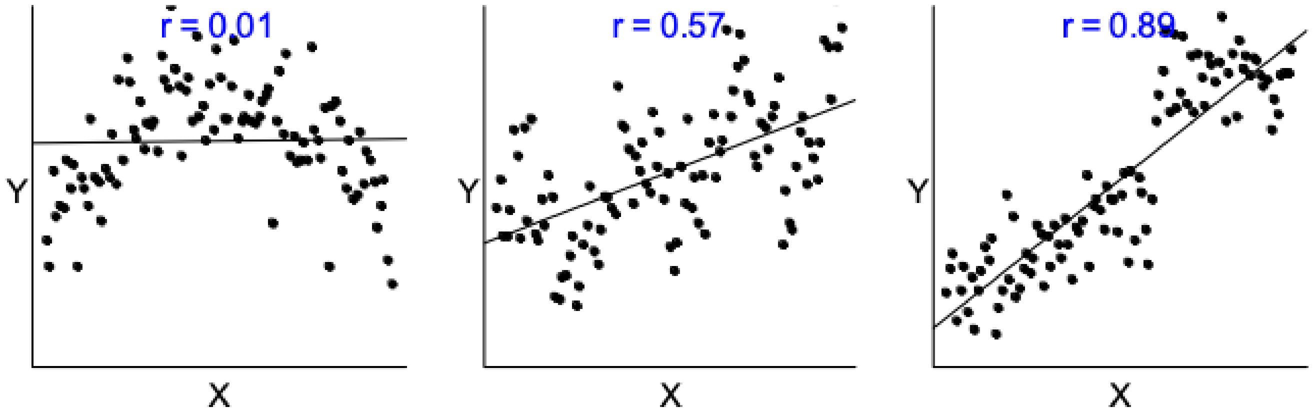
Correlation



Correlation

- The correlation coefficient is appropriate for (almost) linear relationships. Whenever relations deviate from linearity the linear correlation is misleading.

- Examples where the simple correlation coefficient is not really informative:



Correlation & Regression

Linear correlation and linear regression are directly related. The regression slope B can be written as

$$B = \frac{S_{XY}}{S_X^2} = r_{XY}\frac{S_Y}{S_X}$$

Note $r_{XY} = \frac{S_{XY}}{S_X S_Y}$.

If the correlation (covariance) is positive, the slope B is also positive. If the correlation (covariance) is negative, the slope B is also negative.

If both variables X and Y are standardized (i.e., mean values are zero and standard deviations one), then $B = r_{YX}$.

Regression: Predicted Values

We can predict the numeric value of Y for a given value of X using a linear regression. The predicted outcome (i.e., $\hat{Y}$) is a simple linear transformation of X:

$$\hat{Y}_i = A + BX_i$$

The predicted values $\hat{Y}_i$ represent only the part of variation/variance in Y that is 'explained' by the independent variable X.

Prediction: *income ~ age*

Using the estimated regression equation,

$$\hat{\text{Income}}_i = 930.6 + 9.2 \cdot Age_i$$

we can calculate the conditional means for given values of age (these conditional means are typically called predicted or fitted values):

$$\text{Age} = 20 : \hat{\text{Income}}_{\text{Age}=20} = 930.6 + 9.2 \cdot 20 = 1115$$
$$\text{Age} = 40 : \hat{\text{Income}}_{\text{Age}=40} = 930.6 + 9.2 \cdot 40 = 1229$$
$$\text{Age} = 65 : \hat{\text{Income}}_{\text{Age}=65} = 930.6 + 9.2 \cdot 65 = 1529$$

All the predicted values are on the regression line (they form the regression line). If the *linearity* assumption does not hold, the predicted values might be a poor estimate of the 'real' and observed (if observable) conditional means.

How to Run LM in R?

- use regression function `lm()`

```
out.lm <- lm(income ~ age, data = incex)
summary(out.lm)    # prints summary stats of the fitted regression moc
```

How to Run LM in R?

- use regression function `lm()`

```
summary(out.lm)   # prints summary stats of the fitted regression moc
```

```
##
## Call:
## lm(formula = income ~ age, data = incex)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -687.86 -163.92    2.97  155.25  709.12
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 930.6422    18.6254   49.97   <2e-16 ***
## age           9.1753     0.4287   21.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 229.7 on 1920 degrees of freedom
## Multiple R-squared:  0.1926,    Adjusted R-squared:  0.1922
## F-statistic:   458 on 1 and 1920 DF,  p-value: < 2.2e-16
```

Scatterplot with Linear Regression Lines

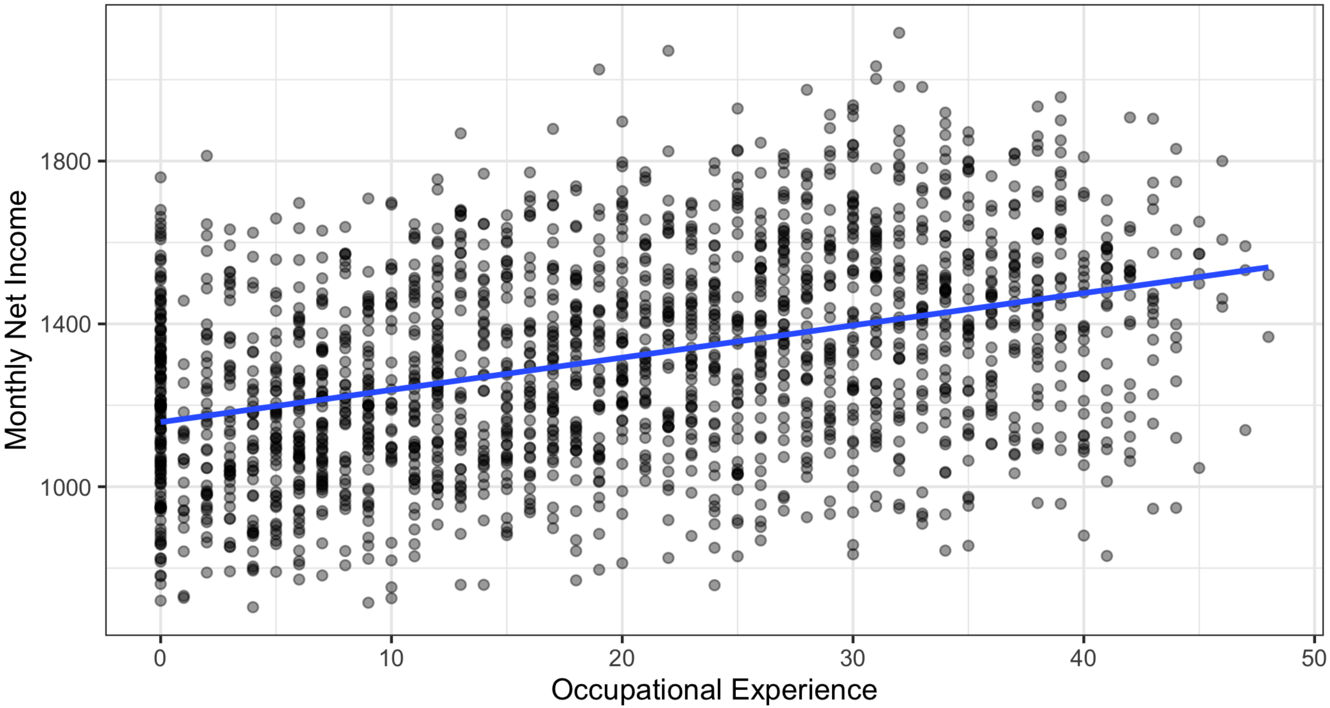
- Use `abline()` to add a linear regression line.

```
plot(income ~ age, data = incex, cex = .6, main = 'income ~ age') # p
abline(out.lm, col = 'blue', lwd = 2)  # use `lm` object
```

Scatterplot with Linear Regression Lines

- Use `geom_smooth()` to add a linear regression line.

```
ggplot(incex, aes(x=oexp, y=income)) + geom_point(alpha=0.4) + labs(x
        geom_smooth(method='lm', formula= y~x, se = FALSE) + theme_bw
```



How to get predicted values in R?

- use the `predict()` function.

```
out.lm <- lm(income ~ age, data = incex)
predict(out.lm, data.frame(age = c(20, 40, 65)))
```

```
##        1        2        3
## 1114.148 1297.654 1527.037
```

Scatterplots with Predicted Values

- Use `points()` to add predicted values.

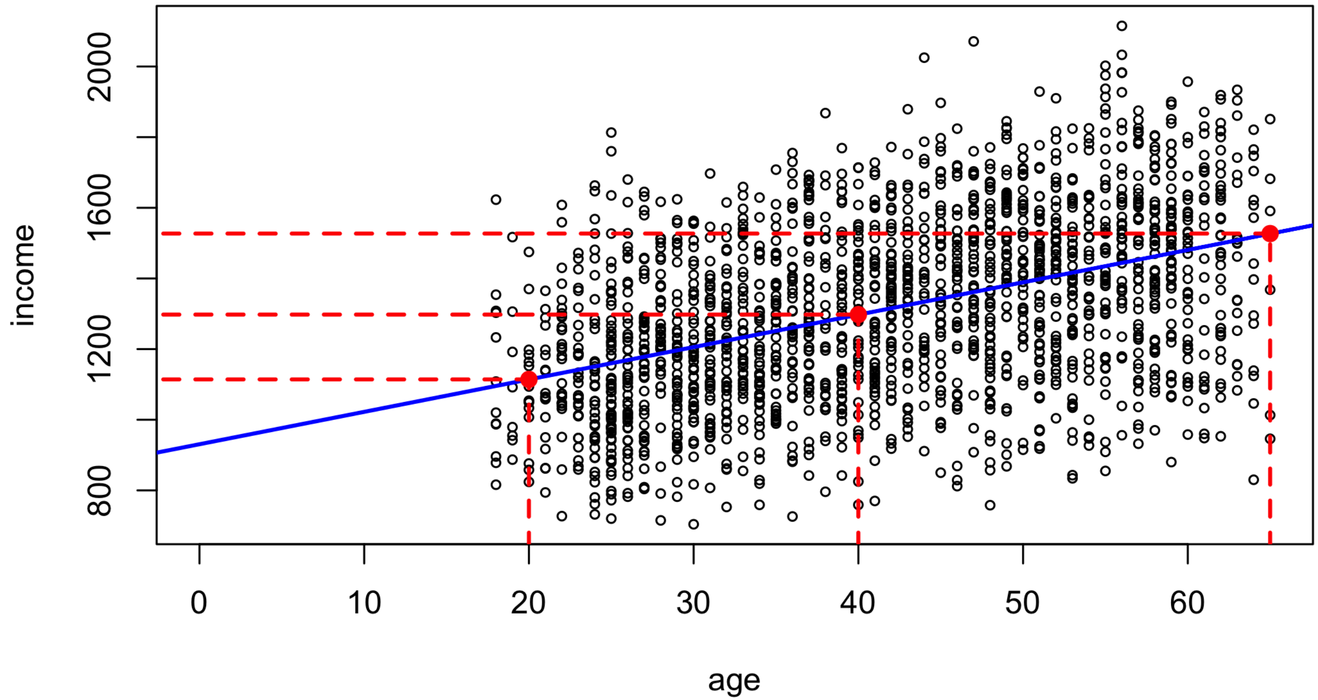
```
# ::::: plot data & predicted values :::::
par(mar = c(4, 4, 1, 1))
plot(income ~ age, data = incex, cex = .6, xlim = c(0, 65)) # plot sc
abline(out.lm, col = 'blue', lwd = 2)                        # add reg

age0 <- c(20, 40, 65)
pre <- predict(out.lm, data.frame(age = age0))               # predict
points(age0, pre, pch = 16, cex = 1.2, col = 'red')          # add po

# add lines for predicted values
segments(age0, rep(0, 3), age0, pre, col = 'red', lwd = 2, lty = 2)
segments(rep(-10, 3), pre, age0, pre, col = 'red', lwd = 2, lty = 2)
```

Scatterplots with Predicted Values

- Use `points()` to add predicted values.



Scatterplots with Predicted Values

- Use `geom_points()` to add predicted values.

```
pred.dat <- data.frame(age=age0, pred=pre)

ggplot(incex, aes(x=age, y=income)) + geom_point() + labs(x = 'Occupa
        geom_smooth(method='lm', formula= y~x, se = FALSE) +
        geom_point(dat = pred.dat, mapping = aes(x=age, y=pred), col=
        theme_bw()
```