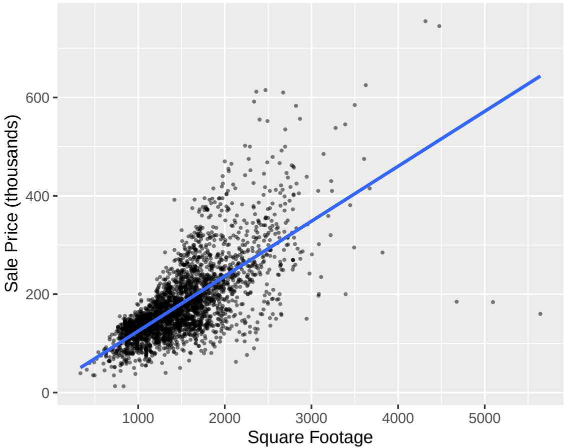
Lecture 2
Simple Linear Regression
STAT 452, Spring 2021

- Linear regression is a useful and widely applied approach to supervised learning.
- It is important to have a good understanding of linear regression before studying more complex statistical learning methods.
- Many fancy statistical learning approaches can be seen as generalizations or extensions of linear regression.

Simple linear regression is a method for fitting a straight line to data that show a linear trend when displayed on a scatterplot. It is a useful tool for making predictions for a quantitative response variable.



Simple Linear Regression Model

Let $\{(x_i, y_i) : i = 1, \cdots, n\}$ be a collection of n data points. A **simple linear regression model** expressing the relationship between y_i and x_i is given by:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- y_i response variable
- x_i predictor variable
- β_0 intercept parameter
- β_1 slope parameter
- ϵ_i is the random error term; assume $\epsilon_i \sim N(0, \sigma^2)$

It is called "simple" linear regression because there is only one predictor variable.

Terminology

In statistical / machine learning we can use the following terms interchangeably:

x: predictor variable, explanatory variable, independent variable, input variable, feature

y: response variable, dependent variable, target variable, output variable

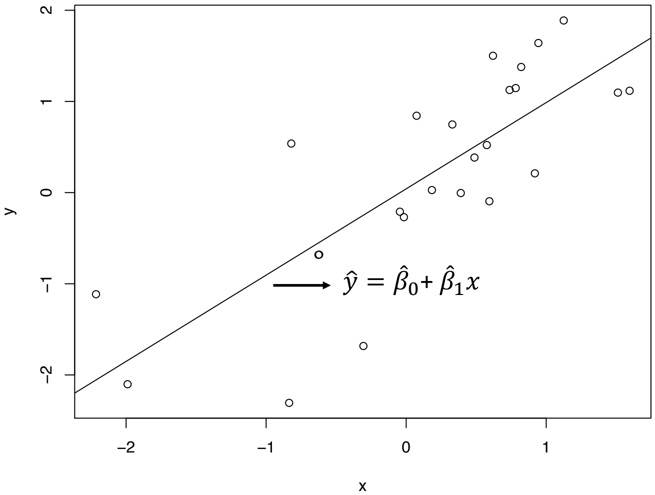
β_0, β_1: regression parameters or coefficients

Fitted Values and Residuals

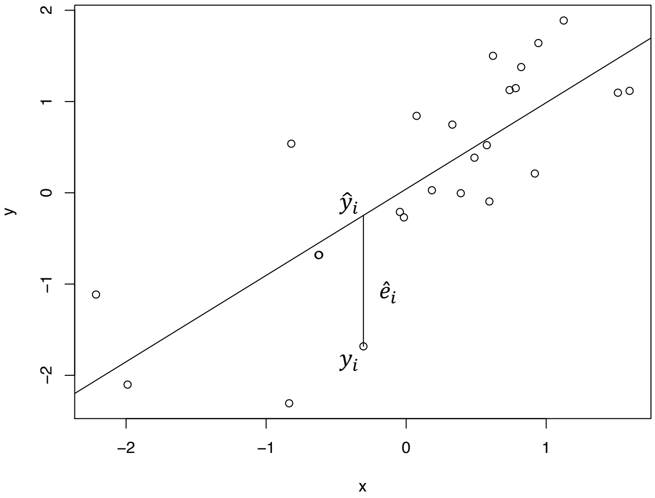
- The line that we we estimate, or fit to the data in the scatterplot, is written as
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$
where $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the estimates of the unknown regression parameters β_0 and β_1.

- The fitted (or predicted) value for the i^{th} observation (x_i, y_i):
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

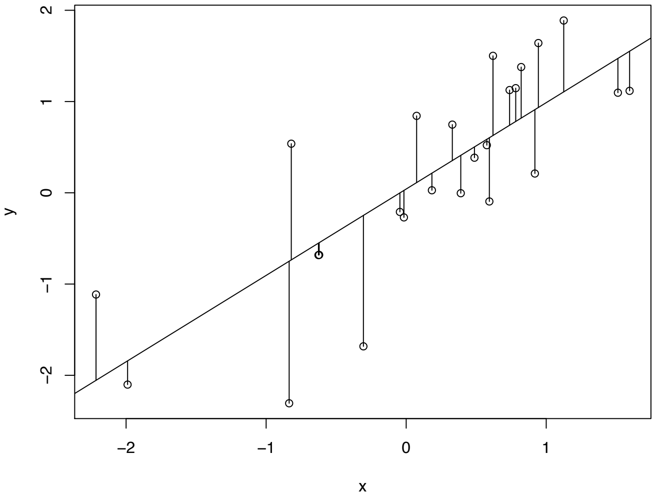
- The **residual** for the i^{th} observation is the difference between the observed value (y_i) and the predicted value $(\hat{y}_i)$:
$$\hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$



$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
y
x
2
1
0
−1
−2
−2
−1
0
1



2
1
0
−1
−2
y
−2
−1
0
1
x
$\hat{y}_i$
$\hat{e}_i$
y_i


2
1
0
−1
−2
y
−2
−1
0
1
x

Sum of Squared Residuals

- Intuitively, a line that fits the data well has small residuals.
- The **least squares line** minimizes the **residual sum of squares**:

$$RSS = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- That is, out of all possible lines we could draw on the scatterplot, the least squares line is the "best fit" since it has the smallest sum of squared residuals.

Least Squares Estimates

Using some calculus, one can show that the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the residual sum of squares (RSS) are given by:

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r\frac{s_y}{s_x}$$

where $\bar{x}$ and $\bar{y}$ are the sample means, s_x and s_y are the sample standard deviations, and r is the correlation coefficient.

Interpretation

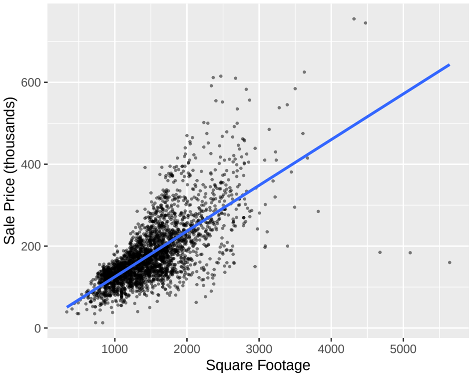
- **Slope**: an increase in the explanatory variable (x) by one unit is associated with a change of $\hat{\beta}_1$ in the predicted response ($\hat{y}$).
- **Intercept**: the prediction for the response variable ($\hat{y}$) when the value for the explanatory variable is zero ($x = 0$). It may not make sense to try to interpret the intercept depending on the application.

Example

- Data set on residential properties in Ames, Iowa from 2006 to 2010, which can be accessed though the R package `AmesHousing`.
- The data set contains 2930 observations (properties) and 81 variables.
- For this example, we fit a simple linear regression model with sale price (`Sale_Price`) as the response variable, and total above ground living space in square feet (`Gr_Liv_Area`) as the predictor.
- To read more about this data package in the R help menu type `help(make_ames)` and `help(ames_raw)`

```
> library(tidyverse)
> library(AmesHousing)
> ames <- make_ames()

> ggplot(ames, aes(x = Gr_Liv_Area, y = Sale_Price / 1000)) +
    geom_point(size = 0.5, alpha = 0.5) +
    geom_smooth(method = "lm", se = FALSE) +
    labs(x = "Square Footage", y = "Sale Price (thousands)")
```



```
> lm1 <- lm(Sale_Price ~ Gr_Liv_Area, data = ames)

> summary(lm1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13289.634   3269.703   4.064 4.94e-05 ***
Gr_Liv_Area   111.694      2.066  54.061  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 56520 on 2928 degrees of freedom
Multiple R-squared:  0.4995,Adjusted R-squared:  0.4994
F-statistic:  2923 on 1 and 2928 DF,  p-value: < 2.2e-16

> coef(lm1) # just extract coefficients
(Intercept) Gr_Liv_Area
  13289.634     111.694
```

Example

(a) Write the equation for the least squares regression line.

(b) Interpret the slope of the model.

(c) What is the predicted sales price for a property with 2000 square feet of above ground living area?

R code to make prediction:

```
> predict(lm1, newdata = data.frame(Gr_Liv_Area = 2000))
       1
236677.6
```

(d) Use R to predict the sales price for a property with 4500 square feet of above ground living area? Based on the scatterplot would you have much confidence in this prediction?

Coefficient of Determination

The **coefficient of determination** (R^2) is a measure of how well the linear regression model fits the data.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares (total variability in the response variable)
- $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the residual sum of squares (unexplained variability)

Coefficient of Determination

- R^2 can be interpreted as the proportion of variability in the response variable y that is explained by x.
- $0 \leq R^2 \leq 1$; the closer R^2 is to 1, the better the linear regression model fits the data.
- R^2 can be computed as the correlation coefficient r squared.
- R^2 is arguably one of the most commonly misused statistics. Always look at a scatterplot of your data first, and check whether fitting a line makes sense and for any outliers.

Example

- Based on the summary output $R^2 = 0.4995$ (see `Multiple R-squared`). Therefore, about 50% of the variability in sale price can be explained by above ground living area.

- Alternatively, we can compute R^2 by taking the sample correlation (using the `cor()` function) and then squaring it.

```
> cor(ames$Sale_Price, ames$Gr_Liv_Area)^2
[1] 0.4995379
```