STA130H1F

Class #9 - Linear Regression

Prof. Nathalie Moon

2018-11-19

Last class

Supervised learning (prediction) for categorical outcome: classification trees

- interpretation
- constructing trees
- prediction
- ROC

This Class

→ continuous y. outcome

Supervised learning for numerical outcome: Linear regression

- Relationship between two continuous variables
- Linear regression models
  - Estimation
  - Interpretation
- Prediction
- Assessing model fit

Relationships between two variables

Advertising Example

- Suppose that we are statistical consultants hired by a client to provide advice on how to improve sales of a particular product.
- The `Advertising` data set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper.

```
glimpse(Advertising)
```

```
## Observations: 200
## Variables: 4
## $ TV        <dbl> 230.1, 44.5, 17.2, 151.5, 180.8, 8.7, 57.5, 120.2, 8...
## $ radio     <dbl> 37.8, 39.3, 45.9, 41.3, 10.8, 48.9, 32.8, 19.6, 2.1,...
## $ newspaper <dbl> 69.2, 45.1, 69.3, 58.5, 58.4, 75.0, 23.5, 11.6, 1.0,...
## $ sales     <dbl> 22.1, 10.4, 9.3, 18.5, 12.9, 7.2, 11.8, 13.2, 4.8, 1...
```

outcome

Advertising Example

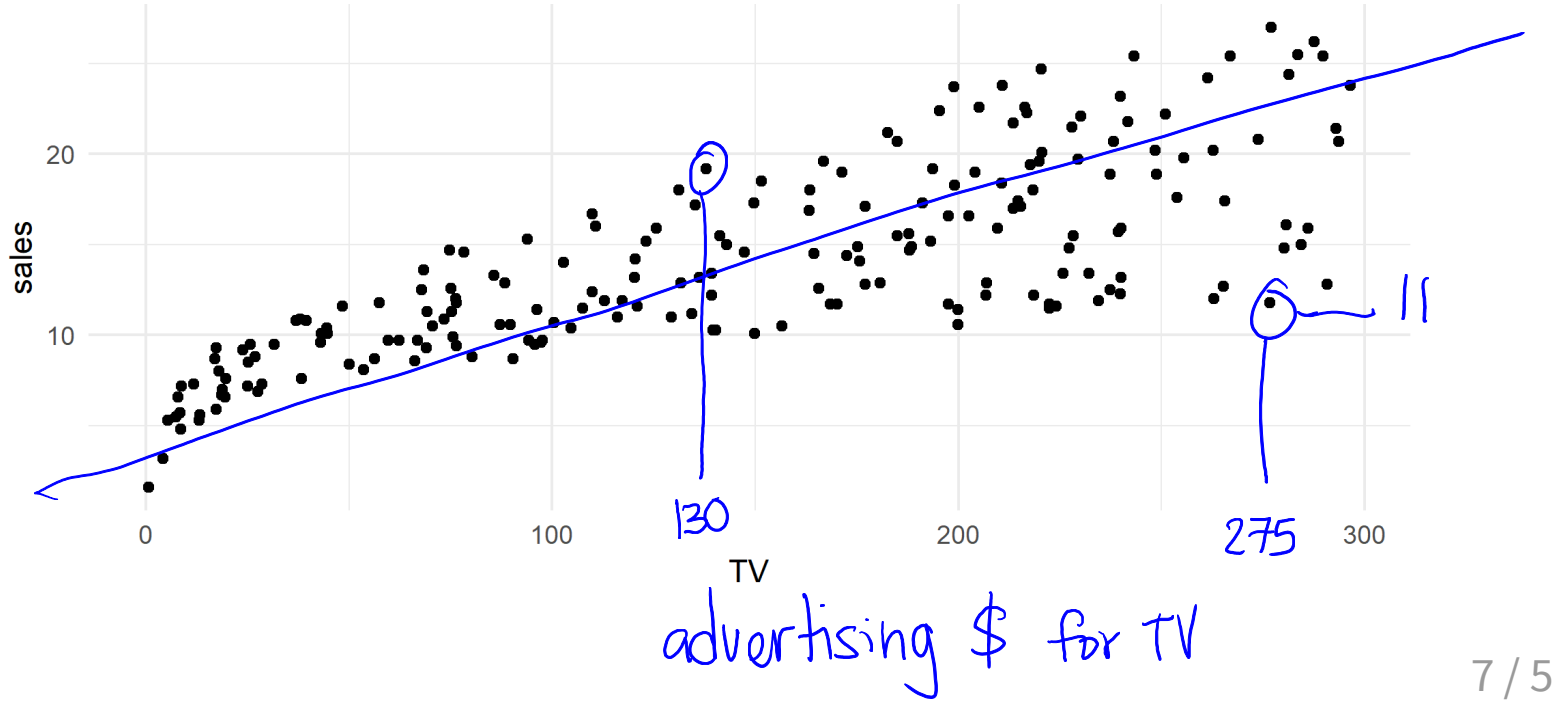
- It is not possible for our client to directly increase sales of the product, but they can control the advertising expenditure in each of the three media.
- Therefore, if we determine that there is an association between advertising and sales, then we can instruct our client to adjust advertising budgets, thereby indirectly increasing sales.

We can control the $ spent on advertising
↳ we hope this will indirectly affect sales

Increasing sales through advertising

What is the relationship between `sales` and `TV` budget?

```
Advertising %>% ggplot(aes(x = TV, y = sales)) + geom_point() +
theme_minimal()
```



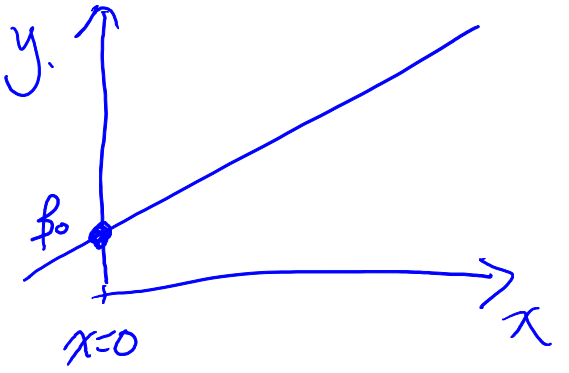
Increasing sales through advertising

- In general, as the budget for `TV` increases `sales` increases.
- Although, sometimes increasing the `TV` budget didn't increase `sales`.
- The relationship between these two variables is approximately linear.

Linear Relationships

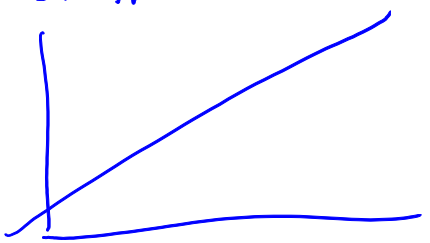
A perfect linear relationship between an independent variable x and dependent variable y has the mathematical form:

$$y = \beta_0 + \beta_1 x.$$

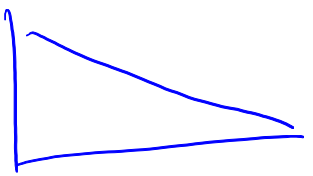


- β_0 is called the y-intercept
  - Why? Bcs when $x=0$, $y=\beta_0$
- β_1 is called the slope.

↳ slope of the line

↳ 

→ slope is positive

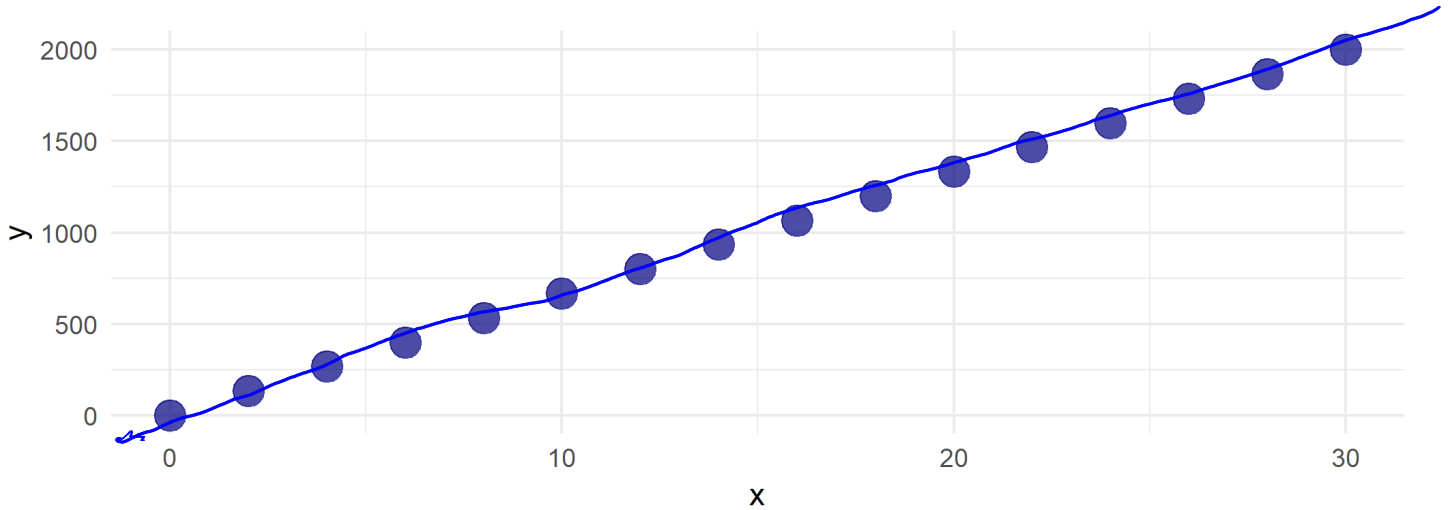


→ slope is negative.

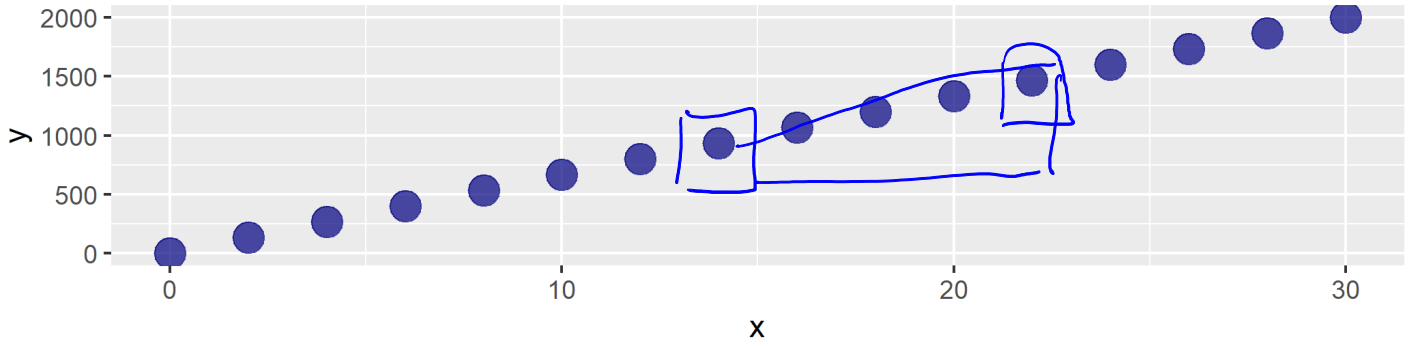
Linear Relationships: The equation of a straight line

Linear Relationships: The equation of a straight line

If the relationship between y and x is perfectly linear, the scatter plot would look like:



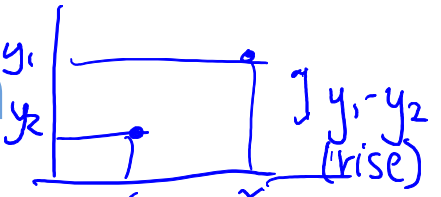
What is the equation of straight line that fits these points?




```
## First four observations
head(data_frame(x = seq(0,30, by = 2),
                y =seq(0, 2000, length.out = length(x))),
     n = 4)
```

```
## # A tibble: 4 x 2
##       x     y
##   <dbl> <dbl>
## 1     0    0
## 2     2  133.
## 3     4  267.
## 4     6  400
```

Fitting a straight line to data



Use analytic geometry to find the equation of the straight line: pick any two points $(x^{(1)}, y^{(1)})$ and $(x^{(2)}, y^{(2)})$ on the line.

$x_1 - x_2$ (run).

The slope is:

$$m = \frac{y^{(1)} - y^{(2)}}{x^{(1)} - x^{(2)}}.$$ → slope of the line.

So the equation of the line with slope m passing through $(x^{(1)}, y^{(1)})$ is

$$y - y^{(1)} = m(x - x^{(1)}) \Rightarrow \boxed{y = mx + b,}$$

β_1 β_0

✓

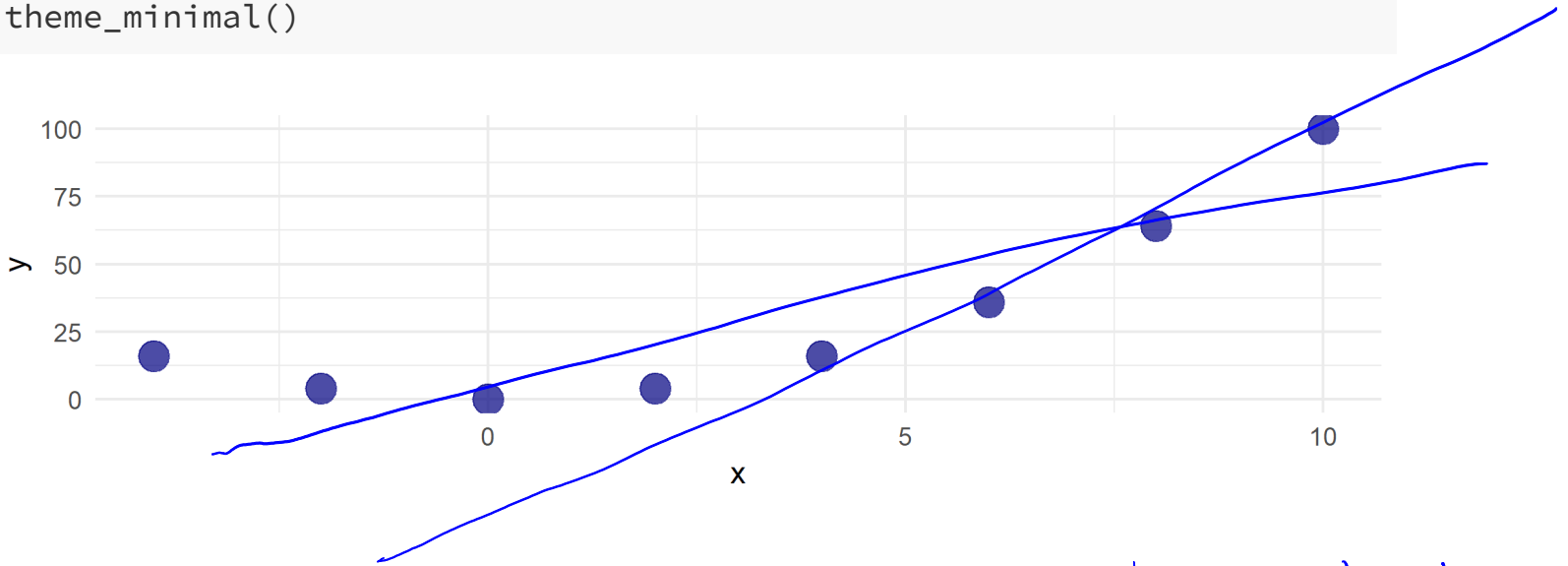
where $b = y^{(1)} - mx^{(1)}$.

plug in the numbers to get the intercept.

Linear Relationships: The equation of a straight line

What is the equation of the 'best' straight line that fits these points?

```
data_frame(x = seq(-4,10, by = 2), y = x^2) %>%
ggplot(aes(x,y)) +
geom_point(cex = 5, colour = "navyblue", alpha = 0.7) +
theme_minimal()
```



some data can't be summarized well by a straight line.

Relationships between two variables

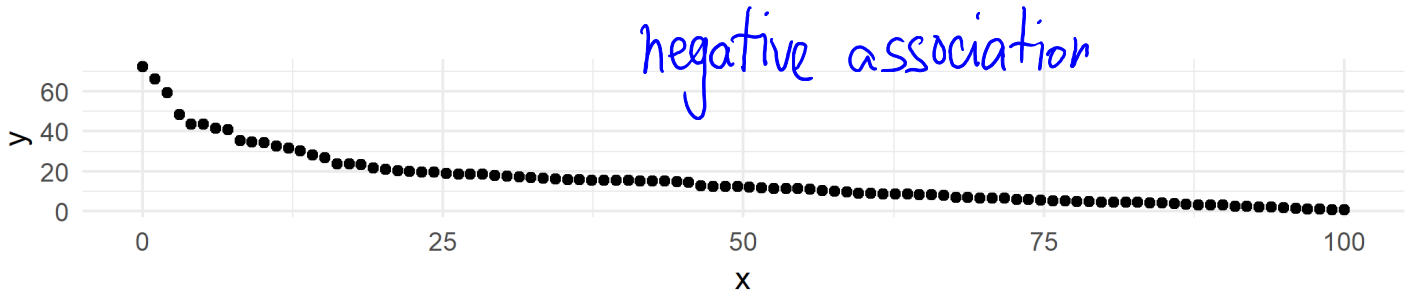
Relationships between two variables

- Sometimes the relationship between two variables in non-linear.
- If the relationship is non-linear then fitting a straight line to the data is not useful in describing the relationship.

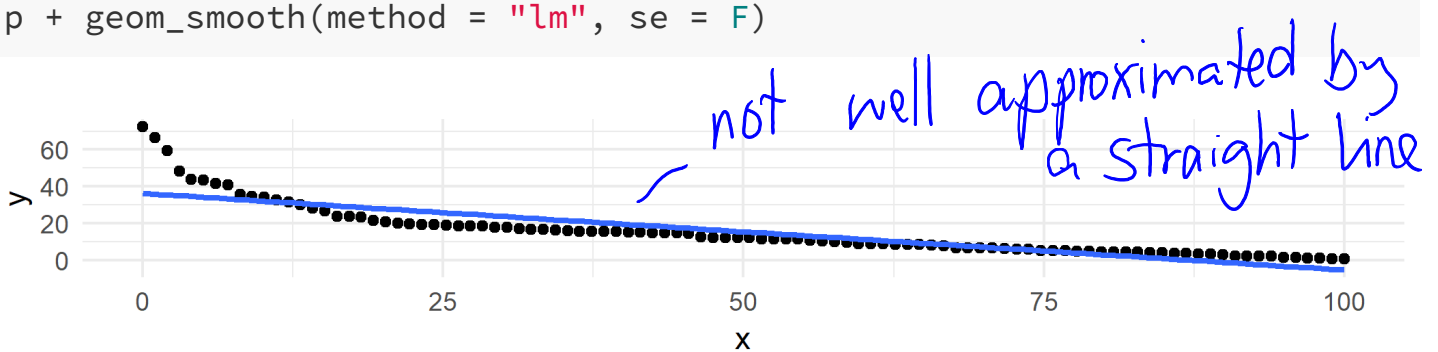
Example of Non-linear relationships

- Let y be life expectancy of a component, and x the age of the component.
- There is a relationship between y and x, but it is not linear.

```
p <- data_frame(x = age, y = life_exp) %>%
ggplot(aes(x = x, y = y)) + geom_point() + theme_minimal()
p
```




```
p + geom_smooth(method = "lm", se = F)
```



Tidy the Advertising Data

- Each market is an observation, but each column is the amount spent on TV, radio, newspaper advertising.

```
head(Advertising, n=3)
```

```
## # A tibble: 3 x 4
##      TV radio newspaper sales
##   <dbl> <dbl>     <dbl> <dbl>
## 1 230.   37.8      69.2  22.1
## 2  44.5  39.3      45.1  10.4
## 3  17.2  45.9      69.3   9.3
```

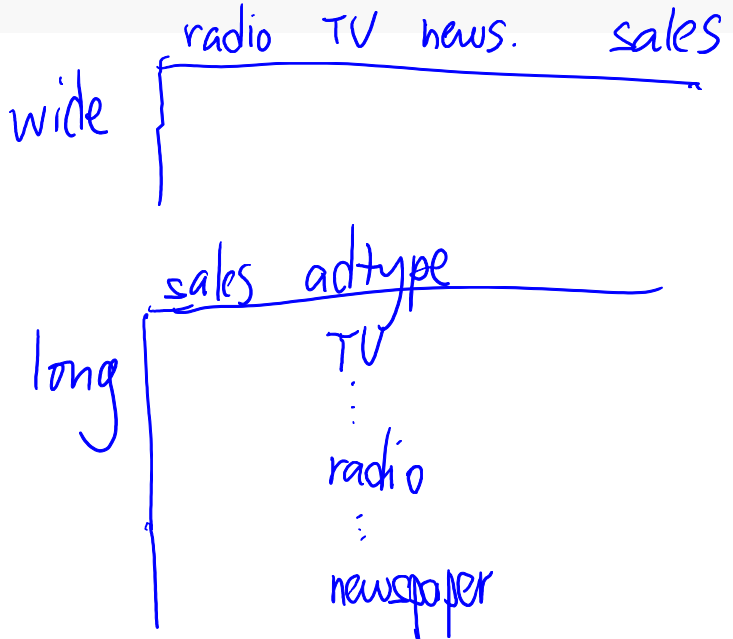
- The data are not tidy since each column corresponds to the values of advertising budget for different media.

Tidy the Advertising Data

- Tidy the data by creating a column for advertising budeget and another column for type of advertising.
- We can use the `gather` function in the `tidyr` library (part of the `tidyverse` library) to tidy the data.

```
Advertising_long <- Advertising %>%
select(TV, radio, newspaper, sales) %>%
gather(key = adtype, value = amount, TV, radio, newspaper)
head(Advertising_long)
```

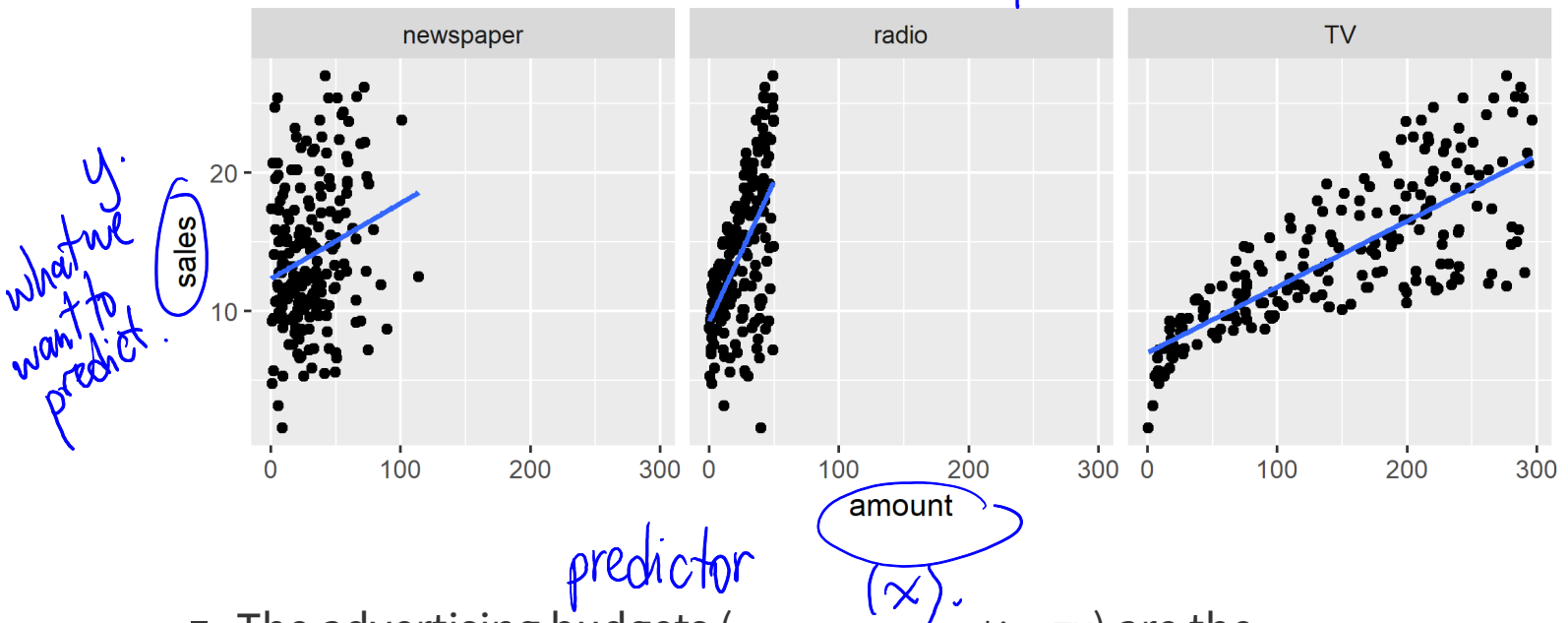
```
## # A tibble: 6 x 3
##   sales adtype amount
##   <dbl> <chr>   <dbl>
## 1  22.1 TV      230.
## 2  10.4 TV       44.5
## 3   9.3 TV       17.2
## 4  18.5 TV      152.
## 5  12.9 TV      181.
## 6   7.2 TV        8.7
```



Advertising Data

```
Advertising_long %>%
ggplot(aes(amount, sales)) +
geom_point() +
geom_smooth(method = "lm", se = FALSE) +
facet_grid(. ~ adtype)
```

- approx. linear.
- positive



what we want to predict. y.

predictor (x).

- The advertising budgets (`newspaper`, `radio`, `TV`) are the input/independent/covariates and the dependent variable is sales.

Linear Regression Models

Simple Linear Regression

The simple linear regression model can describe the relationship between sales and amont spent on radio advertising through the model

eq. of a straight line

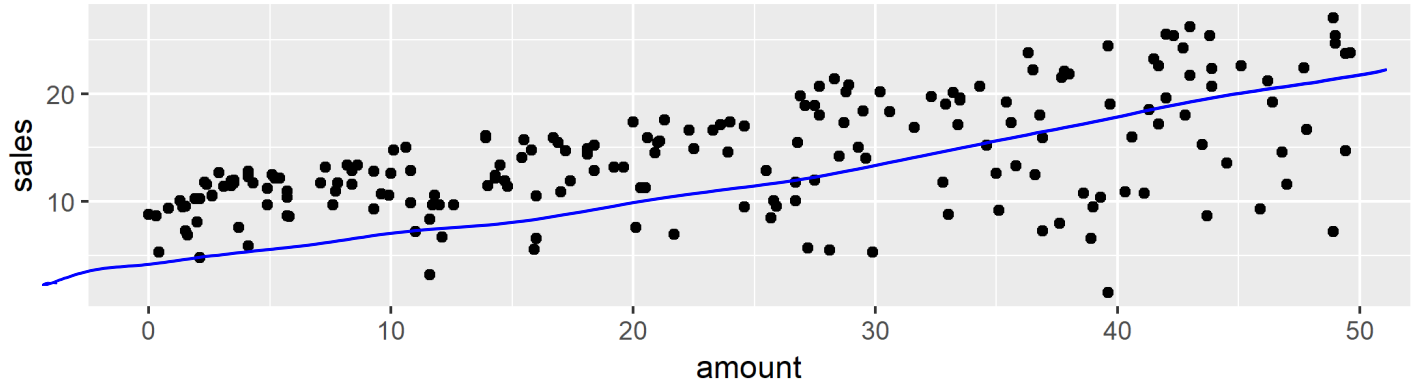
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

epsilon-i → error (random).

where $i = 1, \ldots, n$ and n is the number of observations.

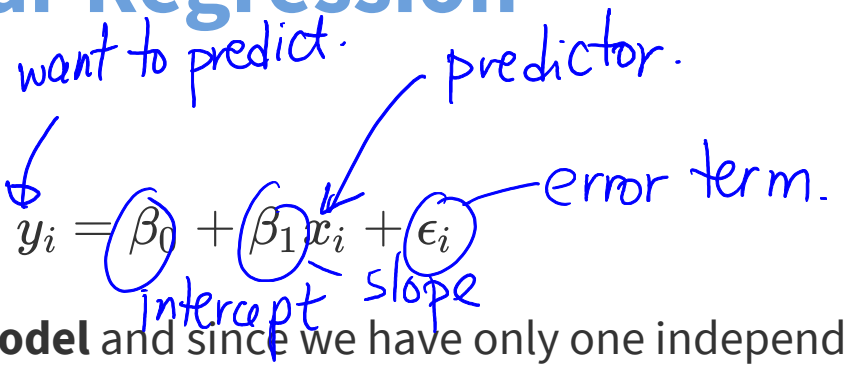
```
Advertising_long %>%
filter(adtype == "radio") %>%
ggplot(aes(amount, sales)) +
geom_point()
```

Q: What is the "best" line to describe these data.



Simple Linear Regression

The equation:



is called a **regression model** and since we have only one independent variable it is called a *simple regression model*.

- y_i is called the dependent or target variable. outcome
- β_0 is the intercept parameter.
- x_i is the independent variable, covariate, feature, or input.
- β_1 is called the slope parameter.
- ϵ_i is called the error parameter.

Multiple Linear Regression

In general, models of the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

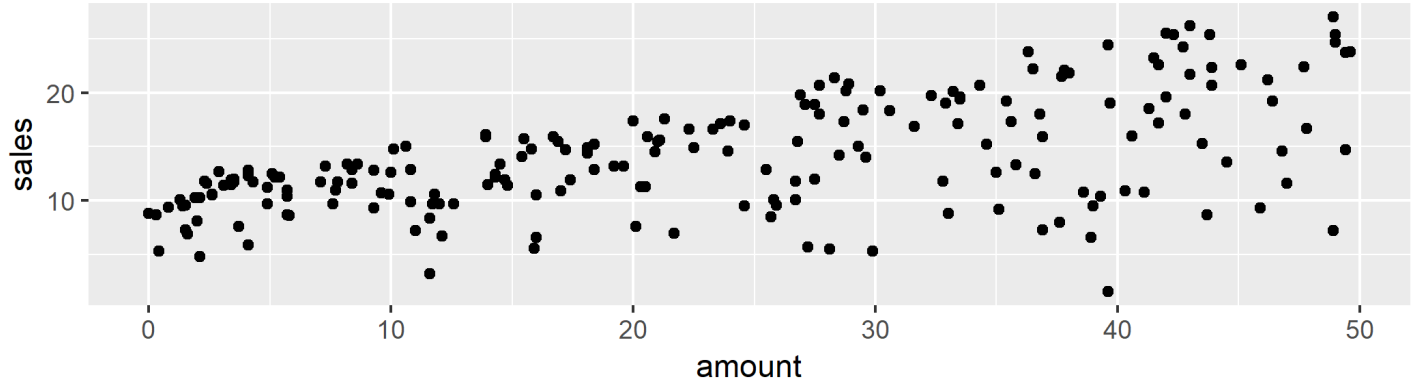
error term

where $i = 1, \ldots, n$, with $k > 1$ independent variables are called *multiple regression models*.

- The β_j's are called parameters and the ϵ_i's errors.
- The values of of neither β_j's nor ϵ_i's can ever be known, but they can be estimated.
  ↳ bcs they are parameters
- The "linear" in Linear Regression means that the equation is linear in the parameters β_j. → no functions of the β parameters.
- This is a linear regression model: $y_i = \beta_0 + \beta_1 \sqrt{x_{i1}} + \beta_2 x_{i2}^2 + \epsilon_i$ $\quad x_{i3} = \sqrt{x_{i1}}$
- This is not a linear regression model (i.e., a nonlinear regression model): $y_i = \beta_0 + \sin(\beta_1) x_{i1} + \beta_2 x_{i2} + \epsilon_i$
  ↳ makes it not a linear model.

Least Squares

Fitting a straight line to Sales and Radio Advertising



y x

```
## # A tibble: 6 x 2
##   sales amount
##   <dbl>  <dbl>
## 1  22.1   37.8 ①
## 2  10.4   39.3 ②
## 3   9.3   45.9
## 4  18.5   41.3
## 5  12.9   10.8
## 6   7.2   48.9
```

$m = \frac{y_1 - y_2}{x_1 - x_2} = \frac{22.1 - 10.4}{37.8 - 39.3} = -7.8$?

$b =$

Fitting a straight line to Sales and Radio Advertising

```
## # A tibble: 6 x 2
##   sales amount
##   <dbl>  <dbl>
## 1  22.1   37.8
## 2  10.4   39.3
## 3   9.3   45.9
## 4  18.5   41.3
## 5  12.9   10.8
## 6   7.2   48.9
```

Fitting a straight line to Sales and Radio Advertising

```
## # A tibble: 6 x 2
##   sales amount
##   <dbl>  <dbl>
## 1  22.1   37.8
## 2  10.4   39.3
## 3   9.3   45.9
## 4  18.5   41.3
## 5  12.9   10.8
## 6   7.2   48.9
```

$m = \frac{22.1-10.4}{37.8-39.8} =$ -5.85, negative! - but that doesn't make sense to describe our data.

$b = 22.1 - \frac{22.1-10.4}{37.8-39.8} \times 37.8 =$ 243.23.

So, the equation of the straight line is:

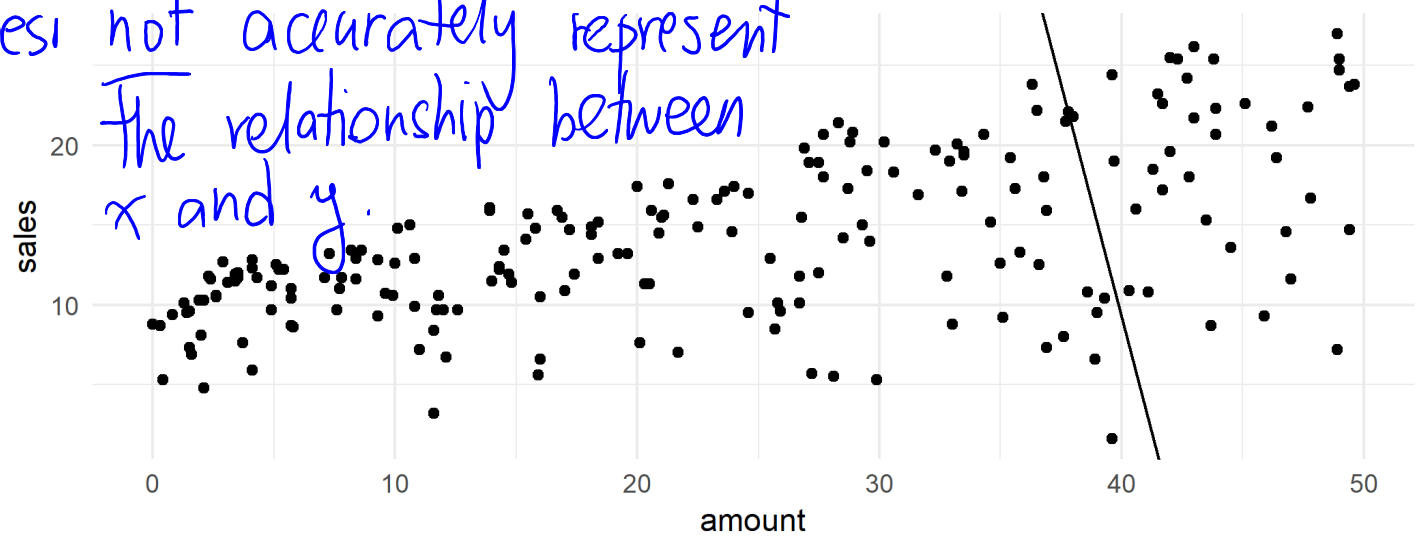
$$y = 243.23 - 5.85x.$$ - check the math.

Fitting a straight line to Sales and Radio Advertising

The equation $y = 243.23 - 5.85x$ is shown on the scatter plot.

```
Advertising_long %>%
filter(adtype == "radio") %>%
ggplot(aes(amount, sales)) +
geom_point() +
geom_abline(intercept = 243.23, slope = -5.85) + theme_minimal()
```

does not accurately represent the relationship between x and y.

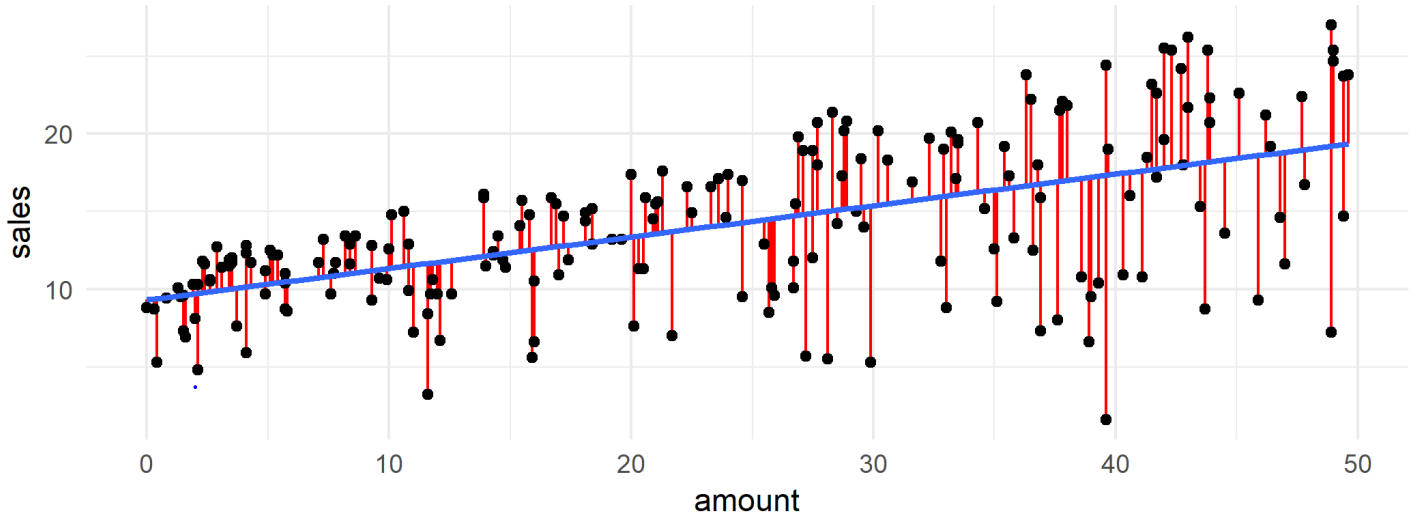


Fitting a straight line to Sales and Radio Advertising

- For a fixed value of `amount` spent on radio ads the corresponding `sales` has variation. It's neither strictly increasing nor decreasing.
- But, the overall pattern displayed in the scatterplot shows that *on average* `sales` increase as `amount` spent on radio ads increases.

Least Squares

The Least Squares approach is to find the y-intercept β_0 and slope β_1 of the straight line that is closest to as many of the points as possible.



Interactive applet

Estimating the coefficients: Least Squares

To find the values of β_0 and slope β_1 that fit the data best we can minimize the sum of squared errors $\sum_{i=1}^{n} \epsilon_i^2$:

# of observations. points

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$-(\beta_0 + \beta_1 x_i)$ → lies on the line $y = \beta_0 + \beta_1 x$

So, we want to minimize a function of β_0, β_1

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2,$$

→ not random and are known for our data

where x_i's are numbers and therfore constants.

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

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$\Rightarrow$ find $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize $L(\beta_0, \beta_1)$.

($\hat{\beta}_0$, $\hat{\beta}_1$ are estimates.)

$$\frac{\partial L}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1) = 0$$

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum x_i = 0.$$

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} y_i - \hat{\beta}_1 \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)$$

$$= \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{, where } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$\frac{\partial L}{\partial \beta_1} = 0$$

Estimating the coefficients: Least Squares

- The derivative of $L(\beta_0, \beta_1)$ with respect to β_0 treats β_1 as a constant. This is also called the partial derivative and is denoted as $\frac{\partial L}{\partial \beta_0}$.

- To find the values of β_0 and β_1 that minimize $L(\beta_0, \beta_1)$ we set the partial derivatives to zero and solve:

$$\frac{\partial L}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0,$$

$$\frac{\partial L}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0.$$

⇒ 2 equations and 2 unknowns.

The values of β_0 and β_1 that are solutions to the above equations are denoted $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively.

Estimating the coefficients: Least Squares

It can be shown that

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{(\sum_{i=1}^{n} y_i x_i) - n\bar{x}\bar{y}}{(\sum_{i=1}^{n} x_i^2) - n\bar{x}^2},$$

formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$

where, $\bar{y} = \sum_{i=1}^{n} y_i/n$, and $\bar{x} = \sum_{i=1}^{n} x_i/n$.

$\hat{\beta}_0$ and $\hat{\beta}_1$ are called the **least squares estimators** of β_0 and β_1.

Estimating the Coefficients Using R - Formula syntax in R

$y = x$

like the equal sign in model equation

The R syntax for defining relationships between inputs such as amount spent on `newspaper` advertising and outputs such as `sales` is:

$y \sim x$

```
sales ~ newspaper
```

The tilde `~` is used to define the what the output variable (or outcome, on the left-hand side) is and what the input variables (or predictors, on the right-hand side) are.

A formula that has three inputs can be written as

```
sales ~ newspaper + TV + radio
```

one outcome

3 predictors, separated by '+' signs.

Estimating the Coefficients Using `lm()`

`lm()` → linear model.

lm(y ~ x , data)

```
mod_paper <- lm(sales ~ newspaper, data = Advertising)
mod_paper_summary <- summary(mod_paper)
mod_paper_summary$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 12.3514071 0.62142019 19.876096 4.713507e-49
## newspaper    0.0546931 0.01657572  3.299591 1.148196e-03
```

$\hat{\beta}_0$ $\hat{\beta}_1$

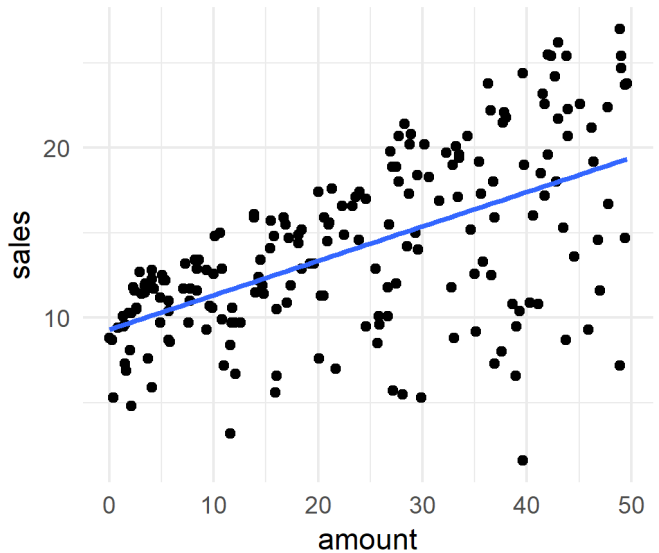
- least squares estimates. → next week

- `(Intercept)` is the estimate of $\hat{\beta}_0$. = 12.3514071
- `newspaper` is the estimate of $\hat{\beta}_1$. ≈ 0.0546931

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

Estimating the Coefficients Using R

```
Advertising_long %>%
  filter(adtype == "radio") %>%
  ggplot(aes(amount, sales)) +
  geom_point() +
  geom_smooth(method = "lm",
              se = FALSE) +
  theme_minimal()
```



- The blue line is the estimated regression line with intercept 12.35 and slope 0.05.
- `geom_smooth(method = "lm", se = FALSE)` adds the linear regression to the scatterplot without a confidence interval for the linear regression line (this is set via `se = FALSE`).

Interpreting the Slope and Intercept with a Continuous Explanatory Variable

The estimated linear regression of `sales` on `newspaper` is:

$$y_i = 12.35 + 0.05x_i,$$

where y_i is sales in the i^{th} market and x_i is the dollar amount spent on newspaper advertising in the i^{th} market.

- The **slope** $\hat{\beta}_1$ is the amount of change in y for a unit change in x.
  - In this example:

- The **intercept** $\hat{\beta}_0$ is the average of y when $x_i = 0$.
  - In this example:

Prediction using a Linear Regression Model

Prediction using a Linear Regression Model

After a linear regression model is estimated from data it can be used to calculate predicted values using the regression equation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

$\hat{y}_i$ is the predicted value of the i^{th} response y_i. (with predictor x_i)

The i^{th} residual is

$$e_i = y_i - \hat{y}_i.$$

true response

predicted response (fitted value, bcs it comes from the fitted regression line)

↳ difference between prediction and true value.

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

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Prediction using a Linear Regression Model

The amount spent on newspaper advertising in the first market is:

```
Advertising %>% filter(row_number() == 1)
```

$\hat{y} = 12.35 + 0.05x$

```
## # A tibble: 1 x 4
##      TV radio newspaper sales
##   <dbl> <dbl>     <dbl> <dbl>
## 1  230.  37.8      69.2  22.1
```

$16.14 = 12.35 + 0.05(69.2)$

x

- The predicted sales using the regression model is: $12.35 + 0.05 \times 69.2 = 16.14$.
- The observed sales for this region is 22.1.
- The **error** or **residual** is $y_1 - \hat{y}_1 = 5.96$.

Prediction using a Linear Regression Model

The predicted and residual values from a regression model can be obtained using the `predict()` and `residual()` functions.

```
mod_paper <- lm(sales ~ newspaper, data = Advertising)
sales_pred <- predict(mod_paper)
head(sales_pred)
```

```
##        1        2        3        4        5        6
## 16.13617 14.81807 16.14164 15.55095 15.54548 16.45339
```

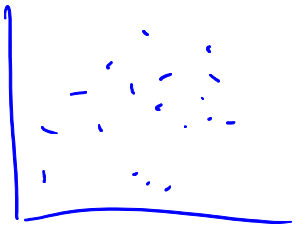
```
sales_resid <- residuals(mod_paper)
head(sales_resid)
```

```
##         1         2         3         4         5         6
##  5.963831 -4.418066 -6.841639  2.949047 -2.645484 -9.253389
```

Measures of Fit for Simple Regression

- The regression model is a good fit when the residuals are small.

Measures of Fit for Simple Regression



- The regression model is a good fit when the residuals are small.
- Thus, we can measure the quality of fit by the sum of squares of the residuals $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

Measures of Fit for Simple Regression

- The regression model is a good fit when the residuals are small.
- Thus, we can measure the quality of fit by the sum of squares of the residuals $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.
- This quantity depends on the units in which y_i's are measured. A measure of fit that does not depend on the units is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

  - R^2 is often called the coefficient of determination.
  - Range of R^2: $(0, 1)$
  - R^2 is the proportion of the variability in y which is explained by the regression model

good model has R^2 close to 1

bad model has R^2 close to 0

Measure of Fit for Simple Regression

The `summary()` method calculates R^2

```
mod_paper <- lm(sales ~ newspaper, data = Advertising)
mod_paper_summ <- summary(mod_paper)
mod_paper_summ$r.squared
```

```
## [1] 0.05212045
```

- $R^2 = 0.0521204$. This indicates a poor fit.

Using Linear Regression as a Machine Learning/Supervised Learning Tool

The `diamonds` data set contains the prices and other attributes of almost 54,000 diamonds. The variables are as follows:

```
library(tidyverse)
library(modelr)
glimpse(diamonds)
```

```
## Observations: 53,940
## Variables: 10
## $ carat   <dbl> 0.23, 0.21, 0.23, 0.29, 0.31, 0.24, 0.24, 0.26, 0.22, ...
## $ cut     <ord> Ideal, Premium, Good, Premium, Good, Very Good, Very G...
## $ color   <ord> E, E, E, I, J, J, I, H, E, H, J, J, F, J, E, E, I, J, ...
## $ clarity <ord> SI2, SI1, VS1, VS2, SI2, VVS2, VVS1, SI1, VS2, VS1, SI...
## $ depth   <dbl> 61.5, 59.8, 56.9, 62.4, 63.3, 62.8, 62.3, 61.9, 65.1, ...
## $ table   <dbl> 55, 61, 65, 58, 58, 57, 57, 55, 61, 61, 55, 56, 61, 54...
## $ price   <int> 326, 326, 327, 334, 335, 336, 336, 337, 337, 338, 339,...
## $ x       <dbl> 3.95, 3.89, 4.05, 4.20, 4.34, 3.94, 3.95, 4.07, 3.87, ...
## $ y       <dbl> 3.98, 3.84, 4.07, 4.23, 4.35, 3.96, 3.98, 4.11, 3.78, ...
## $ z       <dbl> 2.43, 2.31, 2.31, 2.63, 2.75, 2.48, 2.47, 2.53, 2.49, ...
```

Goal: Predict the price of diamonds based on carat size.

Predicting the Price of Diamonds

Let's select training and test sets.

```
set.seed(2)
diamonds_train <- diamonds %>%
mutate(id = row_number()) %>%
sample_frac(size = 0.8)

diamonds_test <- diamonds %>%
mutate(id = row_number()) %>%
# return all rows from diamonds where there are not
# matching values in diamonds_train, keeping just
# columns from diamonds.
anti_join(diamonds_train, by = 'id')
```

Predicting the Price of Diamonds

- Now fit a regression model on `diamonds_train`.

```
mod_train <- lm(price ~ carat, data = diamonds_train)
mod_train_summ <- summary(mod_train)
mod_train_summ$r.squared
```

```
## [1] 0.8488647
```

85% of the variability in price is explained by our model.

- Evaluate the prediction error using root mean square error using the training model on `diamonds_test`.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- RMSE can be used to compare different sizes of data sets on an equal footing and the square root ensures that RMSE is on the same scale as y.

→ "small" ⇒ good predictions
→ "big" ⇒ bad predictions

Predicting the Price of Diamonds using Simple Linear Regression

fit model on training data

- Calculate RMSE using test and training data.

```
y_test <- diamonds_test$price
yhat_test <- predict(mod_train, newdata = diamonds_test)
n_test <- length(diamonds_test$price)

# test RMSE
rmse <- sqrt(sum((y_test - yhat_test)^2) / n_test)
rmse
```

$\hat{y}$ for test data

```
## [1] 1553.295
```

since these are similar, there is no evidence of overfitting.

```
y_train <- diamonds_train$price
yhat_train <- predict(mod_train, newdata = diamonds_train)
n_train <- length(diamonds_train$price)

# train RMSE
sqrt(sum((y_train - yhat_train)^2) / n_train)
```

```
## [1] 1547.391
```

Predicting the Price of Diamonds using Multiple Linear Regression

We will add other variables to the regression model to investigate if we can decrease the prediction error.

```
mrmod_train <- lm(price ~ carat + cut + color + clarity,
                  data = diamonds_train)
mrmod_train_summ <- summary(mrmod_train)
mrmod_train_summ$r.squared
```

```
## [1] 0.9152898
```

91.5% of the variability in price is explained by our model.

```
y_test <- diamonds_test$price
yhat_test <- predict(mrmod_train, newdata = diamonds_test)
n_test <- length(diamonds_test$price)
mr_rmse <- sqrt(sum((y_test - yhat_test)^2) / n_test)
mr_rmse
```

```
## [1] 1149.881
```

= RMSE

- The simple linear regression model had $R^2 = 0.8488647$ and RMSE = 1553.2953095.

⇒ our new model has larger R^2, and also smaller RMSE.

Summary of measures of Fit for Simple Regression

R^2	$RMSE$
$1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
0 to 1	0 to ∞
large value $\Rightarrow$ good fit	small value $\Rightarrow$ good fit
relative measure of fit (no units)	absolute measure of fit (same units as y)
Used to assess model fit (with training data)	Used to assess prediction error (with training or test data)

Overview

Supervised learning for numerical outcome: Linear regression

- Relationship between two continuous variables
- Equation of a straight line
- Linear regression models
  - Estimation
  - Interpretation
- Prediction
- Assessing model fit: R^2 and RMSE