DSA 8010 - simple linear regression

Correlation and regression

The inferential tools covered so far have provided ways to assess association when binary (grouping variables) have been involved.

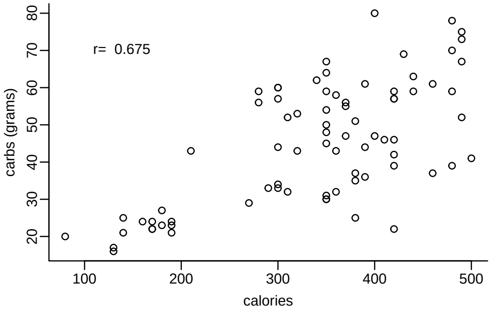
Method	types of variables
inference on two means	numeric & binary
inference on the difference between proportions	binary & binary
simple linear regression	numeric & numeric

Descriptive analysis of two numeric variables

- Make a scatterplot of the two variables.
- Calculate Pearson's correlation (r) to summarize the direction and strength of the linear relationship between the variables.
- If the variables have a highly non-linear relationship, consider calculating Spearman's rank correlation or transforming one or both variables.

Example (Starbucks)

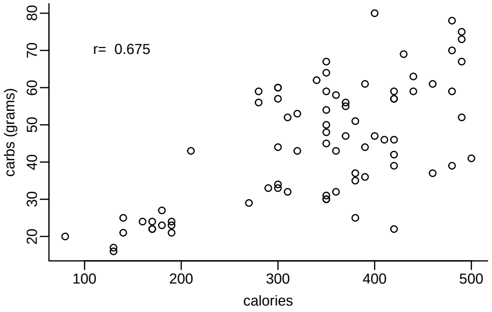
Open Intro Statistics, 4th edition, Diez et al. Each observation in this dataset is one menu item from Starbucks. The calories and carbs (grams) are recorded for each item.



What do the scatterplot and correlation indicate about the relationship between calories and carbs?

Example (Starbucks)

Open Intro Statistics, 4th edition, Diez et al. Each observation in this dataset is one menu item from Starbucks. The calories and carbs (grams) are recorded for each item.



If we observed a new menu item with 250 calories, what would the data predict the carbs to be?

Simple linear regression

The ultimate goal of simple linear regression (SLR): predict a response variable using an explanatory variable.

Examples:

- Use an apple's weight to predict its shelf life.
- Use the height of a certain species of tree to predict its age.
- Use the rate of property crime in a county to predict the rate of violent crime.

Simple linear regression

Data. Two quantitative variables. x is the explanatory variable and y is the response variable, measured on n individuals.

	A	X	Y
1	Department	log(property)	log(violent)
2	Lower Salford Twp Po	6.126432408	3.277145
3	Village Of Port Washin	6.406879986	2.351375
4	Duxbury Police Dept	6.432779308	2.97553
5	Wyckoff Police Dept	6.683986532	3.569533
6	Genoa Twp Police Dep	6.730421264	2.140066
7	Granby Police Dept	6.742880636	2.873565
8	Grosse Ile Twp Police I	6.753204519	2.272126
9	Belle Vernon Boro Pol	6.762960694	5.342813
10	Prospect Heights Polic	6.76768829	4.450853

	item	calories	carb
1	8-Grain Roll	350	67
2	Apple Bran Muffin	350	64
3	Apple Fritter	420	59
4	Banana Nut Loaf	490	75
5	Birthday Cake Mini Doughnut	130	17
6	Blueberry Oat Bar	370	47

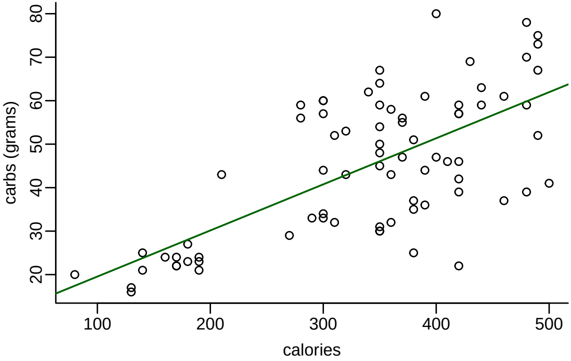
Notation. $(x_i, y_i), i = 1, \ldots, n$ are the pairs of explanatory, response variables.

Statistical model. $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where ϵ_i, $i = 1, \ldots, n$, are i.i.d. and approximately $N(0, \sigma^2)$.

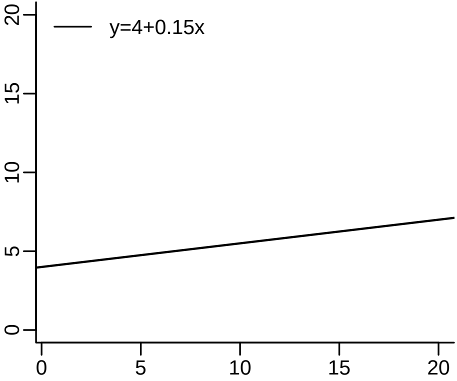
The statistical model describes the relationship between x and y using a straight line.

Simple linear regression

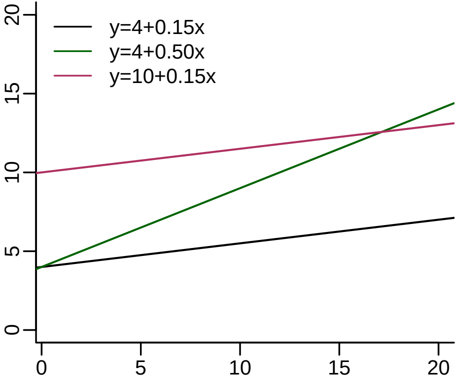
Example (Starbucks)



Equation for a line



Equation for a line



Regression coefficients

Intercept (β_0). y-intercept of the regression line.

Interpretation: expected value (mean) of the response variable (y) when the explanatory variable (x) equals 0.

Slope (β_1). Slope of the line.

Interpretation: the expected increase in the response variable when the explanatory variable increases by 1 unit.

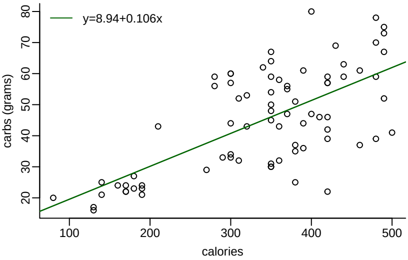
$\beta_1 > 0$: x and y have a positive relationship.
$\beta_1 < 0$: x and y have a negative relationship.
$\beta_1 = 0$: X and Y have no linear relationship.

Statistical model for simple linear regression

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

where ϵ_i, $i = 1, \ldots, n$, are i.i.d. and approximately $N(0, \sigma^2)$.

- The parameters β_0 and β_1 define the regression line.
- The terms $\beta_0 + \beta_1 x_i$ determine the predicted value of the response variable when the explanatory variable is equal to x_i.
- The ϵ_i terms account for represent leftover variability or scatter around the line.



Example (Starbucks)

The estimated regression line for the Starbucks data is

$$y_i = 8.94 + 0.106x_i.$$

- The intercept is equal to 8.94. What is the interpretation of this coefficient?

  A menu item with zero calories is expected to have 8.94 carbs.

- The slope is equal to 0.106. What is the interpretation of this coefficient?

  For every increase of one calorie, the carbs for a Starbucks menu item are expected to increase by 0.106 grams.

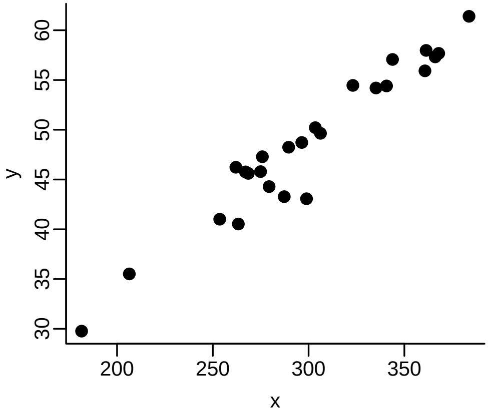
Hat notation

For any parameter in a statistical model, the $\widehat{}$ symbol can be used to generically denote a statistic that is used to estimate that parameter using data.

Examples:

μ: unknown population mean	$\widehat{\mu}$: some estimate of μ calculated from a sample
σ: unknown population standard deviation	$\widehat{\sigma}$: some estimate of σ calculated from a sample.
π: unknown population proportion	$\widehat{\pi}$: some estimate of π calulated from a sample.
β_0, β_1: unknown true intercept and slope	$\widehat{\beta}_0$, $\widehat{\beta}_1$: estimates of the intercept and slope calculated from a sample.

Estimation of regression coefficients



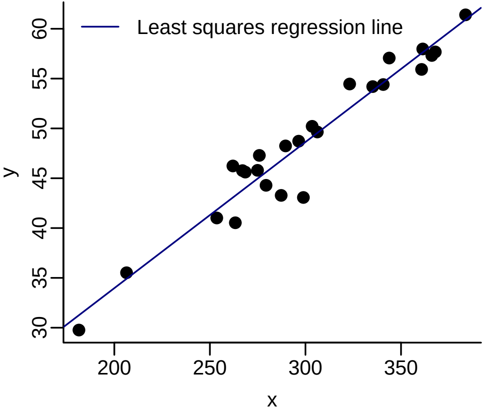
Estimation of regression coefficients

The point estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are found by minimizing the least squares criterion:

$$\sum_{i=1}^{n}(\text{observed } y_i - \text{predicted } y_i)^2 = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

- the quantity $(y_i - (\beta_0 + \beta_1 x_i))^2$ measures the distance between observation i and the regression line defined by β_0 and β_1.
- The least-squares estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ will define a line that is as close to the observed data points as possible.

Estimation of regression coefficients



Formulas for least-squares regression coefficients

The least-squares estimates of β_0 and β_1 can be calculated using the following formulas:

$$\widehat{\beta}_1 = \frac{SXY}{SXX} \qquad \widehat{\beta}_0 = \bar{y} - \widehat{\beta_1}\bar{x}$$

where the SXX and SXY are defined as

$$SXX = \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad SXY = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

In this course, we will use software to calculate the coefficients.

Regression analysis from software

Most statistical software packages present results of a regression in a standard regression table format.

```
Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
-5.403 -1.004  0.407  1.140  3.169

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.639110   2.489935   1.863   0.0753 .
x           0.146647   0.008193  17.899 5.36e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.017 on 23 degrees of freedom
Multiple R-squared:  0.933,	Adjusted R-squared:  0.9301
F-statistic: 320.4 on 1 and 23 DF,  p-value: 5.357e-15
```

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.6391096	2.489935	1.86	0.0753
x	0.1466471	0.008193	17.90	<.0001*

The first row of the table contains $\widehat{\beta}_0$ and the second row contains $\widehat{\beta}_1$.

Prediction in simple linear regression

The prediction equation or equation of the regression line is used to find the predicted value of the response variable ($\widehat{y}_i$) given the value of the explanatory variable (x_i).

The prediction equation is

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

Plug in any value for x_i to get the predicted value of the response variable given that value of the explanatory variable.

Here is the R output using the Starbucks data.

```
Call:
lm(formula = carb ~ calories, data = starbucks)

Residuals:
    Min      1Q  Median      3Q     Max
-31.477  -7.476  -1.029  10.127  28.644

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.94356    4.74600   1.884   0.0634 .
calories     0.10603    0.01338   7.923 1.67e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.29 on 75 degrees of freedom
Multiple R-squared:  0.4556,	Adjusted R-squared:  0.4484
F-statistic: 62.77 on 1 and 75 DF,  p-value: 1.673e-11
```

What are the estimated slope and intercept of the regression line? Interpret their values.

Here is the R output using the Starbucks data.

```
Call:
lm(formula = carb ~ calories, data = starbucks)

Residuals:
    Min      1Q  Median      3Q     Max
-31.477  -7.476  -1.029  10.127  28.644

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.94356    4.74600   1.884   0.0634 .
calories     0.10603    0.01338   7.923 1.67e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.29 on 75 degrees of freedom
Multiple R-squared:  0.4556,    Adjusted R-squared:  0.4484
F-statistic: 62.77 on 1 and 75 DF,  p-value: 1.673e-11
```

What is the predicted carbs for an item with 352 calories?

Inference on regression parameters

Inference on regression parameters

In simple linear regression, we collect a sample of n observations and calculate the estimated coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

- $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are statistics calculated from the data. The true β_0 and β_1 are unknown parameters.

Recap: the standard error of a statistic is the standard deviation of a statistic.

- A different sample from the same population would yield different estimates of β_0 and β_1. The standard error quantifies the variability among these estimates.
- Notation.
  $SE_{\widehat{\beta}_0}$ = standard error of $\widehat{\beta}_0$
  $SE_{\widehat{\beta}_1}$ = standard error of $\widehat{\beta}_1$

Standard errors in regression

Software programs calculate estimated standard errors ($\widehat{SE}_{\widehat{\beta}_0}$, $\widehat{SE}_{\widehat{\beta}_1}$) for each coefficient. These are found in the second column of the regression table.

```
Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max 
-5.403 -1.004  0.407  1.140  3.169 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.639110   2.489935   1.863   0.0753 .  
x           0.146647   0.008193  17.899 5.36e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.017 on 23 degrees of freedom
Multiple R-squared:  0.933,	Adjusted R-squared:  0.9301 
F-statistic: 320.4 on 1 and 23 DF,  p-value: 5.357e-15
```

Generally speaking, we expect the estimated regression coefficients from our data to be no more than 2 or 3 standard errors from the true population parameters.

Confidence intervals for regression coefficients

Confidence interval for the intercept. A $(1-\alpha)\times 100\%$ confidence interval for β_0 is

$$\widehat{\beta}_0 \pm t^*_{n-2,\alpha/2}\widehat{SE}_{\widehat{\beta}_0},$$

where $t^*_{n-2,\alpha/2}$ is the $(1-\alpha/2)\cdot 100$th percentile of the t distribution with df$=n-2$.

Confidence interval for the slope. A $(1-\alpha)\times 100\%$ confidence interval for β_1 is

$$\widehat{\beta}_1 \pm t^*_{n-2,\alpha/2}\widehat{SE}_{\widehat{\beta}_1},$$

where $t^*_{n-2,\alpha/2}$ is the $(1-\alpha/2)\cdot 100$th percentile of the t distribution with df$=n-2$.

Example (Starbucks)

Use R to find a 99% CI for the intercept and a 95% CI for the slope.

Example (expenditures)

Suburban towns often spend a large fraction of their municipal budgets on public safety services. A taxpayers' group felt that very small towns were likely to spend large amounts per person because they have small financial bases. The group obtained data on the per capita expenditure for public safety (`Expen`) of 18 suburban towns in a metropolitan area, as well as the population of each town in units of 1,000 people (`TownPop`). R was used to find a simple linear regression line to predict expenditures using town population (in thousands). Here is the regression table.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 179.2748    11.0936  16.160 2.49e-11 ***
TownPop      -1.3525     0.3039  -4.451 0.000403 ***
---
```

Use the regression table to find a 90% confidence intervals for the slope.

Example (expenditures)

Suburban towns often spend a large fraction of their municipal budgets on public safety services. A taxpayers' group felt that very small towns were likely to spend large amounts per person because they have small financial bases. The group obtained data on the per capita expenditure for public safety (`Expen`) of 18 suburban towns in a metropolitan area, as well as the population of each town in units of 1,000 people (`TownPop`). R was used to find a simple linear regression line to predict expenditures using town population (in thousands). Here is the regression table.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 179.2748    11.0936  16.160 2.49e-11 ***
TownPop      -1.3525     0.3039  -4.451 0.000403 ***
---
```

Predict the expenditures per capita for a town with a population of 54,000.

Hypothesis test for β_0

Hypotheses. $H_0 : \beta_0 = b$; $H_A : \beta_0 \neq b$ $(<, >)$

Test statistic.

$$t_0 = \frac{\widehat{\beta}_0 - b}{\widehat{SE}_{\widehat{\beta}_0}}$$

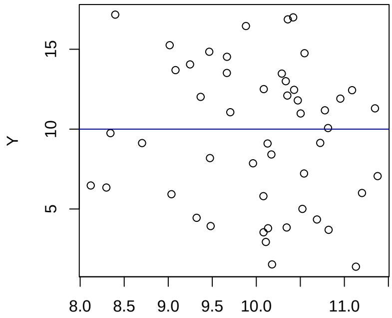
p-value. Use the t distribution with df$=n-2$.
For the two-sided alternative:
`2*pt(abs(t0),df=n-2,lower.tail=FALSE)`.

Decision. Reject H_0 if the p-value is less than α.

Note: If the hypotheses are $H_0 : \beta_0 = 0$; $H_A : \beta_0 \neq 0$, the test statistic and p-value are given in the regression table.

Hypothesis test for β_1

The most common hypothesis test in simple linear regression is a test of $H_0 : \beta_1 = 0$.



If $\beta_1 = 0$, the regression line is flat; i.e., x does not provide any help in making (linear) predictions about y. If $\beta_1 \neq 0$, then there is some benefit to using x to predict y.

Hypothesis test for β_1

Hypotheses. $H_0 : \beta_1 = b$; $H_A : \beta_1 \neq b$ $(<, >)$

Often, $b = 0$.

Test statistic.

$$t_0 = \frac{\widehat{\beta}_1 - b}{\widehat{SE}_{\widehat{\beta}_1}}$$

p-value. Use the t distribution with df$=n-2$.
For the two-sided alternative:
`2*pt(abs(t0),df=n-2,lower.tail=FALSE)`.

Decision. Reject H_0 if the p-value is less than α.

Note: If the hypotheses are $H_0 : \beta_1 = 0$; $H_A : \beta_1 \neq 0$, the test statistic and p-value are given in the regression table.

Example (Starbucks)

Use R to test the following hypotheses. Report the test statistic, p-value, and conclusion.

$H_0 : \beta_1 = 0;\ H_A : \beta_1 \neq 0$

$H_0 : \beta_0 = 6;\ H_A : \beta_0 \neq 6$

Example (expenditures)

Suburban towns often spend a large fraction of their municipal budgets on public safety services. A taxpayers' group felt that very small towns were likely to spend large amounts per person because they have small financial bases. The group obtained data on the per capita expenditure for public safety (`Expen`) of 18 suburban towns in a metropolitan area, as well as the population of each town in units of 1,000 people (`TownPop`). R was used to find a simple linear regression line to predict expenditures using town population (in thousands). Here is the regression table.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 179.2748    11.0936  16.160 2.49e-11 ***
TownPop      -1.3525     0.3039  -4.451 0.000403 ***
---
```

Test using $\alpha = 0.05$ whether the data provide strong evidence in support of the taxpayers' group's claim.