Chapter 1

Introduction to Linear Regression

The need to model the relationship between two numerical variables is among the most ubiquitous, common tasks that arises in almost all disciplines. Linear regression, in its simplest form, is a strategy for finding the best straight line approximation for such relationships.

An illustration

Let us begin with some intuitive explorations on a real-world data set. The file `movies.csv` contains data on a sample of 120 movies produced in the United States, together with information on certain variables associated with each movie. The first few lines of the datafile are shown below

	Movie	USGross	Budget	Stars	Rating	Genre	Run_Time
	<fct>	<dbl>	<dbl>	<dbl>	<fct>	<fct>	<int>
1	White Noise	56.09436	30	2	PG-13	Horror	101
2	Coach Carter	67.26488	45	3	PG-13	Drama	136
3	Elektra	24.40972	65	2	PG-13	Action	100
4	Racing Stripes	49.77252	30	3	PG	Comedy	110
5	Assault on Precinct 13	20.04089	30	3	R	Action	109
6	Are We There Yet?	82.67440	20	2	PG	Comedy	94

The variables include the amount of revenue generated by the movie (`USGross`, in millions of dollars), the budget expended for producing it (`Budget`, in millions of dollars), the movie's star rating (`Stars`), its genre (`Genre`), and its run time (`Run_Time`, in minutes). A question of interest here is whether a movie's box-office success, measured in terms of its `USGross`, is related to the cost of producing it, measured in terms of its `Budget`. Figure 1.1 shows a scatter plot between these two variables. Here we have taken `Budget` as the independent variable, and `USGross` as the dependent variable.

It is relatively straightforward to find the line of best fit for this relationship using linear regression computations. However, it is critically important to recognize that the computed straight line may be a misleading or unreliable model, unless the underlying dataset satisfies

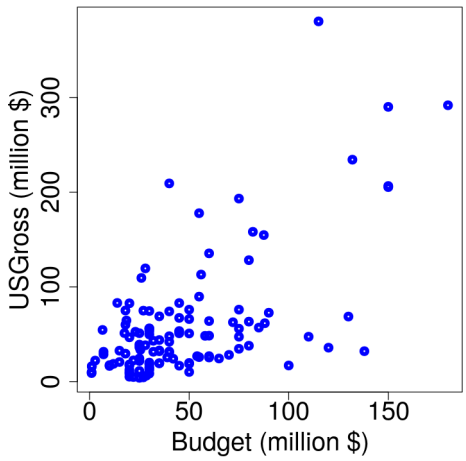


Figure 1.1: Scatterplot of `USGross` vs `Budget` for sample of 120 movies.

certain key theoretical conditions. We will discuss those conditions in detail later. For now, we will just go ahead and find the linear model, and use it to introduce some key regression concepts. The following R code segment reads in the data file and computes the regression line

```
# Read the file "movies.csv" and create a dataframe:
movdat = read.csv(file="https://cs.earlham.edu/~pardhan/sage_
   and_r/movies.csv", header=TRUE, sep=",")
head(movdat)  # Print first few lines of dataframe

# Construct linear regression model using the lm() function:
lout = lm(USGross ~ Budget, data=movdat)
# Print regression output:
summary(lout)
```

The corresponding output is

```
Call:
lm(formula = USGross ~ Budget, data = movdat)

Residuals:
    Min      1Q  Median      3Q     Max
-127.91  -26.55   -3.82   18.18  245.74

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.0315     7.4645   0.942    0.348
Budget        1.1087     0.1271   8.724 2.04e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 49.46 on 118 degrees of freedom
Multiple R-squared:  0.3921,    Adjusted R-squared:  0.3869
F-statistic:  76.1 on 1 and 118 DF,  p-value: 2.041e-14
```

The software output shows the computed slope and intercept in the first column of the table

of `Coefficients`. According to this output, the regression model is

$$\widehat{\texttt{USGross}} = 7.0315 + 1.1087\ \texttt{Budget} \tag{1.1}$$

The left hand side represents the predicted `USGross`. It is customary to use the hat notation to distinguish between predicted value versus true (observed) value of the dependent variable. The slope of the linear model is one of the most practically useful indicators that arises from regression. It gives a precise, quantitative understanding of how a movie's `Budget` affects its predicted `USGross`. In this example, it can be said that for each 1 million dollar increase in a movie's `Budget`, our model predicts its `USGross` increases by 1.1087 million dollars on average.

A common use of regression models is for making predictions. For instance, one might ask how much box-office revenue (`USGross`) is reasonable to expect from a movie whose production budget is \$56 million. If we trust the model, we would simply plug in 56 for `Budget`, and get $\widehat{\texttt{USGross}} = 7.0315 + 1.1087\ (56) = \$69.12$ million. But, there are some subtle points to note about this interpretation:

1. It holds in a statistically average sense. In fact, if the theoretical conditions are satisfied, the predicted values are normally distributed, with mean at \$69.12 million.

2. In practice, we often compute a prediction interval, similar to a confidence interval, that is estimated to contain the true value with some prescribed probability.

Another important use of the model's predictive capability is to assess its overall fitting quality. Suppose, for example, we plug in `Budget` $= 42$, and get

$$\widehat{\texttt{USGross}} = 7.0315 + 1.1087\ (42) = 53.6$$

But, our dataset contains an observed value of `USGross` $= 24.15$ for that same `Budget`. In regression terminology, this error in the model's prediction is known as the residual, defined by

$$\text{residual} = \text{observed value} - \text{predicted value}$$

Since our dataset contains 120 observations, we can compute 120 residual values. There is a very important connection between residuals and the concept of "best-fit," which is at the theoretical core of regression methods:

> In linear regression, the line of best-fit is the one that minimizes the sum of the square of the residuals. This is also known as the least-squares principle.

Returning to our example, Figure 1.2 shows a plot of the residuals for the 120 observations in our data. As expected, some residuals are positive, some are negative, and their mean is 0. How can we use these residuals, or errors, to assess the fitting quality of the model? A common way to do this is to compute what is known as R-squared

$$R^2 = 1\ -\ \frac{\text{variance in residuals}}{\text{variance in observed } \texttt{USGross}} \tag{1.2}$$

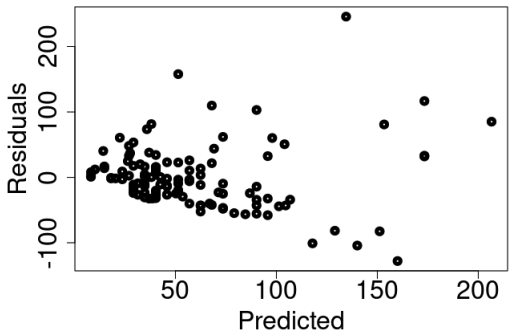


Figure 1.2: Residuals plot for linear regression model (1.1).

The values of R^2 range between 0 and 1. When a model is a very good fit for the data, the residuals and their variance is near 0, resulting in R^2 values close to 1. Conversely, in a very poor model, the residuals will be almost equal to the observed values, with nearly the same variance, resulting in $R^2 \approx 0$. Essentially, R^2 indicates how much variance in the observed y-values is explained by the regression model. In our example $R^2 = 0.392$, as seen in the regression output. Thus, according to our model, `Budget` accounts for about 39% of the variance in `USGross` seen in the dataset. Is that good enough? Or, not? Unfortunately, there is no simple answer. Acceptable values of R^2 vary enormously based on discipline, application type, and other factors. It is also important to note that a high R^2 value, by itself, does not imply we have a good or reliable model. The only way to ensure validity of the model is to verify that the underlying data and process satisfy all the necessary theoretical conditions.

Regression fundamentals

The goal of this section is to introduce essential concepts and foundational material needed to do hands on regression modeling. We will begin with a brief overview of scatterplots and correlation, followed by a discussion on how to compute the regression line, and how to interpret the results. Theoretical issues about model development, assessment and validation will also be addressed as needed.

Scatterplots and correlation

The first step in almost any kind of regression modeling is to get a visual sense of the relationship between the variables of interest. This is commonly done using x-y scatterplots between pairs of numerical variables. If there are several variables of interest, a common strategy is to make a matrix of scatterplots that shows the relationship between every possible pair. Key things to observe in a scatterplot include

a. Shape or form: Does the scatter pattern look approximately linear, or is it curved?

b. Direction or slope: Positive or negative?

c. Strength: Strong, moderate, or weak – depending on how densely the points are clustered together.

d. Outliers: Are there any points located far from the main scatter pattern?

Each of these observations has a subjective element, and involves making some decisions based on experience and other application-specific considerations. Figure 1.3 shows some example plots for practice, where we assume both axes have roughly equal scale. Plots (b), (d), (e) can be considered approximately linear, while (c), (f) are clearly not. The strength of the relationship appears to be moderate or higher in all the plots, except perhaps (d). Plots (c) and (f) exhibit a strong nonlinear association. Plots (b) and (e) have a point or two that may be considered outliers.

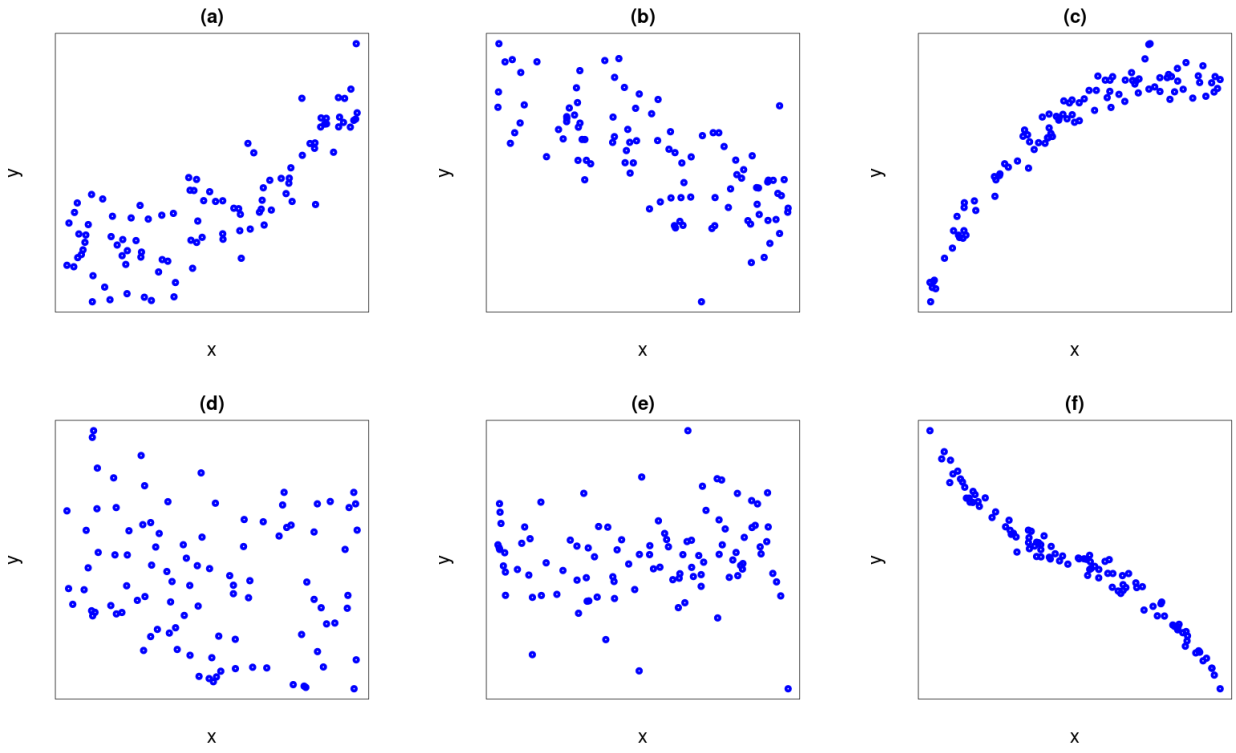


Figure 1.3: Example scatterplots for practicing how to interpret the relationship.

Before proceeding further, it will be helpful to review the following terminology:

- The *independent* variable is also known as the *explanatory* or *predictor* variable. It is customarily plotted on the horizontal axis.

- The *dependent* variable, also known as the *response* or *predicted* variable, is plotted on the vertical axis.

Linear regression methods can be applied to relationships in which the scatterplot can be treated as linear, and has no significant outliers. For such situations, the strength of the linear relationship can be quantified using a special numerical indicator known as the *Pearson correlation* (or simply *correlation*), defined by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\ (y_i - \bar{y})}{s_x\ s_y\ (n-1)} \tag{1.3}$$

where n = number of points in the data set, (x_i, y_i) = explanatory and response variable pair, $\bar{x}$ = mean of x-variable, s_x = standard deviation of x-variable, and similarly for $\bar{y}, s_y$. It is more insightful to rewrite (1.3) in terms of standardized z-scores

$$r = \frac{\sum_{i=1}^{n} z_{x_i}\ z_{y_i}}{n-1} \tag{1.4}$$

where $z_{x_i} = (x_i - \bar{x})/s_x$, and similarly for z_{y_i}.

Although correlations are rarely computed by hand, familiarity with the form of equations (1.3)-(1.4) is insightful. For instance, (1.4) makes it easy to see that r is unitless and, furthermore, its value remains constant even if we reverse the choice of dependent and independent variable. Here are some other important properties of r

a. Its value ranges between -1 and 1. That is, $|r| \leq 1$.

b. Its magnitude is proportional to the strength of linear association: Strong linear association $\Rightarrow |r|$ is close to 1, and weak association $\Rightarrow |r| \sim 0$.

c. The sign of r indicates whether the linear relationship has positive or negative slope.

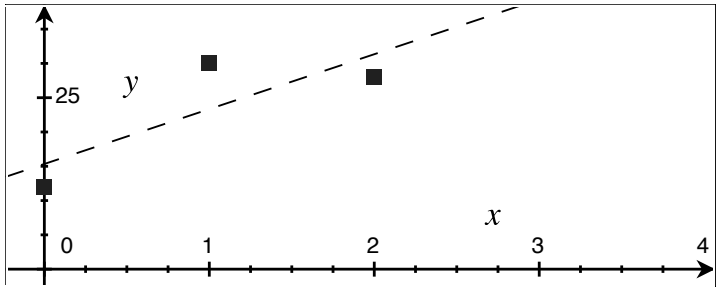
It is extremely important to note that these interpretations only hold for linear relationships. For example, if the underlying relationship is nonlinear, then the value of r has no bearing on the relationship's strength or its direction.

Line of best fit

It is routine practice to compute the line of best fit, or regression line, using software. We will do the same here. However, in the spirit of learning to design and innovate in new directions, we will also briefly study how and why the underlying methods work the way they do. Let us begin with a very simple example that highlights some critically important theoretical ideas.

Example:

x_i	y_i
0	12
1	30
2	28



A very small dataset consisting of only 3 observations is shown in the table, together with a scatter plot of y vs x. Find the best straight line approximation to the plot (see dotted line).

Solution:
The most important aspect of this solution is the overall strategy we use.
Strategy: We will interpret "best straight line" to mean the one that has the least error (measured by the residuals). Here are the key steps we will follow

a. Assume the line is $\hat{y} = mx + b$, where we must find numerical values for m and b.

b. There are 3 data points. For each, find the residual: $e_i = y_i - \hat{y}_i$

c. Compute the sum of the square of the residuals: $f = \sum (e_i)^2$

d. Minimize f: Set $\frac{\partial f}{\partial b} = 0$ and $\frac{\partial f}{\partial m} = 0$, and solve simultaneously for m and b.

Computations based on the above steps:

a. Model: $\hat{y} = mx + b$

b. Residuals: $e_1 = 12 - b$, $e_2 = 30 - m - b$, $e_3 = 28 - 2m - b$

c. Cost function: $f = (12 - b)^2 + (30 - m - b)^2 + (28 - 2m - b)^2$

d. Minimize:
$\frac{\partial f}{\partial b} = -2(12 - b) - 2(30 - m - b) - 2(28 - 2m - b) = 0 \quad \Rightarrow 3b + 3m = 70$
$\frac{\partial f}{\partial m} = -2(30 - m - b) - 4(28 - 2m - b) = 0 \quad \Rightarrow 3b + 5m = 86$
Solve for m, b and get: $m = 8$, $b = 46/3$

Answer: The best straight line approximation is $\hat{y} = 8x + 46/3$

This example illustrates one of the major theoretical ideas at the foundation of regression methods: The least squares principle. Regression models, essentially, begin with a template of the type of curve we want to fit (e.g., linear, polynomial, exponential, etc.). The parameter values in our template are then found by minimizing the sum of the square of the residuals.

Let us now consider the problem of finding the best straight line approximation to any general data set. In theory, the process is the same as that used in the simple example above. However, the implementation is much more convenient if we develop general formulas to compute the slope and intercept of the line.

In standard notation, the regression line is generally written in the form

$$\hat{y} = b_0 + b_1 x \tag{1.5}$$

where b_0 and b_1 are to be found using the least squares process. It can be shown (as we will do in the exercises) that the corresponding line of best fit has slope

$$b_1 = r\frac{s_y}{s_x} \tag{1.6}$$

where r = correlation coefficient, and s_x, s_y = standard deviations of the x, y variables, respectively. Thus b_1 can be readily computed for any given data set. To find b_0 we use the fact that the best straight line must pass through $(\bar{x}, \bar{y})$, the respective mean values of the x, y variables. After computing b_1, we can solve for b_0 by plugging $(\bar{x}, \bar{y})$ into (1.5).

It is common practice in regression analysis to compute and report the value of R^2 defined in (1.2). It turns out this R^2 value is numerically the same as r^2, the square of the correlation coefficient, which makes the computation straightforward. Regression results are typically summarized by including the equation of the straight line, together with a descriptive interpretation of the slope and the R^2 value. We will show exactly how this is done in the examples later.

Assumptions and conditions

Linear regression is, essentially, a strategy for modeling the relationship between two variables in a statistically average sense. It is useful to keep in mind that the data on which the regression model is built is just a sample drawn from a larger population. Ultimately, it is this population we are really interested in, and not the sample that was used to build the model.

There are four conditions that must be satisfied for linear regression between a pair of numerical variables to produce a reliable predictive model

1. Linearity: The relationship between the variables must be approximately linear, with no significant outliers.

2. Normal residuals: The residuals resulting from the model must be approximately normally distributed, with mean=0.

3. Constant variance of residuals: The spread of the data points must remain roughly the same as we move along the regression line.

4. Independent observations: The data points in the sample must consist of independent observations.

Typically, the first 3 conditions are checked graphically, as we will shortly see in the examples. Checking the fourth condition (independent observations) usually requires information about how the data were collected. Ideally, we want a random sample that is representative of the underlying population of interest.

Examples and discussion

One of the interesting features of linear regression is that the computations involved are extremely mundane and straightforward. Perhaps for that exact reason, there is a pervasive tendency to misuse and misinterpret regression models, which can lead to perilously flawed predictions and analyses. In our view, the overwhelming majority of effort spent on learning regression modeling should be devoted to conceptual foundations and theoretical understanding. Accordingly, we will strive to focus on such issues in most of the examples that follow.

Example 1
Let us revisit the movies data set for which we computed a regression model earlier. The data consisted of a sample of 120 movies produced in the United States, together with information on some variables associated with each movie. The variables included: `USGross`, `Budget`, `Stars`, `Genre`, and `Run_Time`. The question we were asking is whether a movie's box-office success, measured in terms of its `USGross`, is related to the cost of producing it, measured in terms of its `Budget`. To this end, we computed the following linear regression model

$$\widehat{\texttt{USGross}} = 7.0315 + 1.1087\,\texttt{Budget}$$

The goal of this example is to address the following questions:

a. Is this a reasonable model?

b. Is a movie's box-office success related to the cost of producing it?

Solution:

a. To determine whether the model is reasonable, we must check whether it satisfies the assumptions and conditions for linear regression.

   The scatterplot in Fig. 1.1 may pass off as roughly linear, though there is at least one high outlier (with `USGross` $\sim$ 400), and possibly also outliers on the low end, with `Budget` values between 100-150.

   Taking a look at Fig. 1.4, we observe that the distribution of residuals is not quite normal, particularly for higher values. But, it may be okay to treat it as nearly normal, especially since the sample size is not small. However, the scatterplot in the leftmost graph shows a significant increase in variability as we move to the right. This is a more serious violation of the requirement of constant variance.

   The final condition we must check is whether the data consists of independent observations. To do this, we require more information about how the data were compiled – e.g., whether the sample is random, or at least representative of movies produced in the United States. Since we do not have access to this information, we cannot assume independence.

   To summarize, our model fails to satisfy two of the required conditions: (1) constant variance of residuals, and (2) independent observations. Thus, it is not a reasonable model for predicting `USGross`, based on `Budget`.

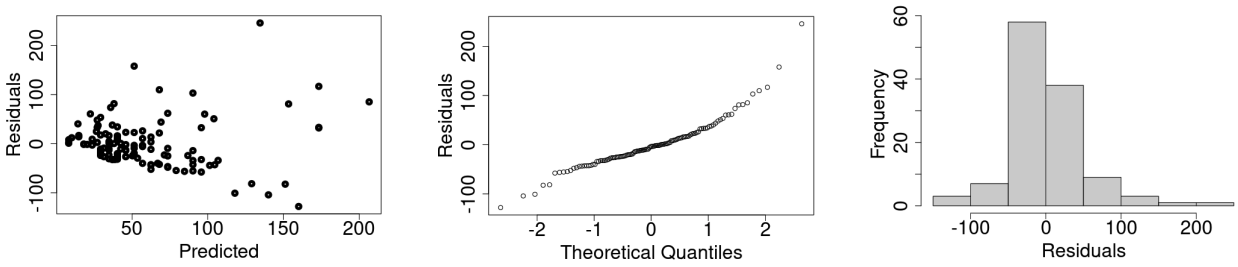
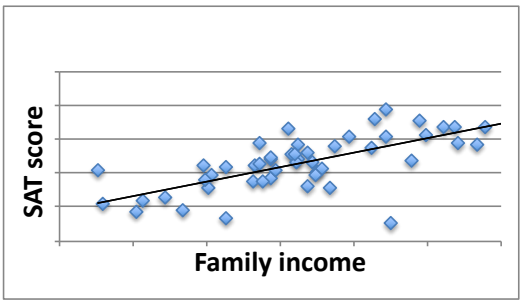


Figure 1.4: Scatterplot, normal probability plot, and histogram of the residuals.

b. This is a somewhat tricky and subtle question. On the one hand, it is important to recognize that a regression model cannot possibly establish a cause-effect relationship, even if the model is perfect in every way. Thus, the model cannot tell us if a movie's box-office success is related to the cost of producing it. On the other hand, if the model had satisfied all the necessary conditions, it could tell us that **on average** higher budget movies tend to be associated with higher box-office revenues. But, since our model does not satisfy the conditions, it cannot even make that claim.

Example 2

The Scholastic Aptitude Test (SAT) is a standardized exam taken by high school students in the U.S. for admission into college. Education research indicates that students from wealthier families tend to have higher SAT scores than those from poorer families. Given below are summary statistics on SAT score and family income for a random sample of 129 high school seniors.



	Mean	SD
SAT score (no units)	1220	109
Family income (in 1000 dollars)	75.8	9.2

A scatter plot of the sampled data suggests the relationship is approximately linear, and the correlation is $r = 0.65$. We want to construct a linear regression model to predict SAT scores from family income.

a. Identify the explanatory variable and the response variable.

b. To the extent possible, check whether the conditions for linear regression are met.

c. Find the equation of the regression line.

d. Find the R^2 value, and explain what it means in statistical terms.

e. Interpret what the slope indicates in this context.

Solution:

a. Explanatory variable = Family income. Response variable = SAT score.
Reason: Since we want to predict SAT scores, it is the response variable.

b. The scatterplot of SAT score vs Family income appears close enough to linear. However, there is one point with a high family income and low SAT score that may be an outlier.

It is hard to tell whether the distribution of residuals is normal. But the sample size is large, and the normality condition is less critical for large samples.

Except for the outlier, the scatter pattern seems to be roughly even, above and below the regression line. Thus, the constant variance condition may be met.

Since the sample is random and large, it is reasonable to treat the observations as independent.

To summarize, except for the single outlier, it appears that the conditions for linear regression are met.

c. Since we want to predict SAT score from family income:
x-variable = family income (in 1000 \$)
y-variable = SAT score (no units)

Given summary statistics are:

$\bar{x} = 75.8, s_x = 9.2, \bar{y} = 1220, s_y = 109, r = 0.65$

Model looks like: $\hat{y} = b_0 + b_1 x$.

$$b_1 = r\frac{s_y}{s_x} = 0.65\left(\frac{109}{9.2}\right) \approx 7.7 \text{ per 1000 dollars}$$

Then we have: $\hat{y} = b_0 + 7.7x$.

Find b_0 by plugging in $(\bar{x}, \bar{y})$:

$1220 = b_0 + 7.7(75.8) \Rightarrow b_0 = 1220 - 7.7 \times 75.8 = 636.34$

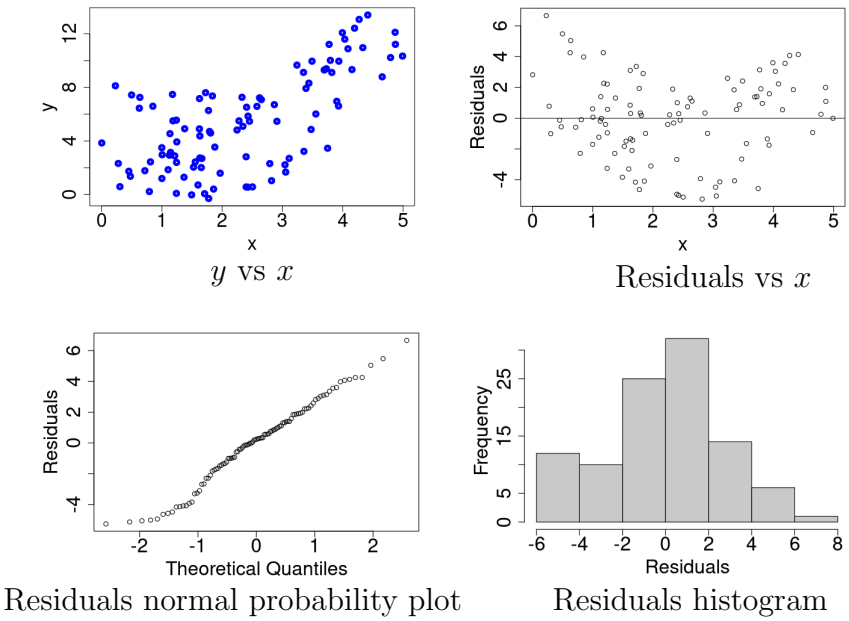
Equation of regression line is:

$\hat{y} = 636.34 + 7.7\,x$ OR $\boxed{\widehat{\text{SAT score}} = 636.34 + 7.7 \text{ (Family income in 1000 \$)}}$

d. Since $r = 0.65$, $R^2 = (0.65)^2 = 0.4225$ OR 42.25%
The R^2 value of 42.25% means that about 42% of the variance in SAT scores is explained by the variance in family income.

e. The slope indicates that for every 1000 dollars increase in a student's family income, the model predicts the student's SAT score increases by about 7.7 points, on average.

Example 3
A dataset contains 100 randomized observations of a pair of x, y numerical variables. Determine whether a linear regression model would be appropriate for predicting y from x, based on the diagnostic plots shown below.



y vs x | Residuals vs x

Residuals normal probability plot | Residuals histogram

Figure 1.5: Diagnostic plots for example 3.

Solution

Linearity: The plot of y vs x appears somewhat nonlinear, particularly at the lower end of the domain. In fact, the overall nonlinearity is easier to see in the plot of Residuals vs x.
Normal residuals: The normal probability plot and histogram of residuals are not perfect, but they may be within range of acceptable, particularly with a sample size of 100.
Constant variance of residuals: The nonlinear shape of the Residuals vs x plot makes it unlikely that the variance will remain close to a constant.
Independent observations: Since the data consists of randomized observations, it is reasonable to assume independence.

To summarize, it would NOT be appropriate to use linear regression for modeling this relationship. The data fails to meet the linearity and the constant variance conditions.

Example 4

The data file `fastfoods.csv` contains nutrition information on a representative sample of 118 food items served at popular fast food chains, including McDonald's, Burger King, Wendy's, etc. For each food item, data is provided about several variables, some of which include its type, serving size (grams), calories (no units), total fat (grams), saturated fat (grams), protein (grams), and sodium (milligrams). Can we predict protein from the amount of calories in fast food? Carry out a linear regression analysis to find out.
Note: An important goal of this example is to illustrate the use of R code and builtin utilities for linear regression.

Solution

The predictor variable is calories, and response variable is protein. We will use the following strategy for regression analysis

a. Assess linearity by plotting protein vs calories.

b. Compute the equation of the regression line.

c. Plot residual diagnostics to check normality and constant variance.

d. Check for independent observations.

e. If all conditions are met, we will summarize the model, and interpret the slope and R^2.

Tasks a-c are implemented in the following R code segment, with output shown later.

```
# Read datafile and store data in a dataframe
fooddat = read.csv(file="https://cs.earlham.edu/~pardhan/sage_
   and_r/fastfoods.csv", header=TRUE, sep=",")

# View the first few lines, and the names of the variables
head(fooddat)
names(fooddat)

# Plot protein vs calories
plot(Protein_g ~ Calories, xlab="Calories", ylab="Protein␣(g)",
    data=fooddat)

# Compute linear model and summarize results
foodout = lm(Protein_g ~ Calories, data=fooddat)
summary(foodout)

# Plot residuals vs fitted values, and plot normal probability
plot(foodout$residuals ~ foodout$fitted, xlab="Predicted", ylab
   ="Residuals")
abline( 0, 0 )    # This plots horizonal reference line at y=0
qqnorm(foodout$residuals, ylab="Residuals")
```

The plot of protein vs calories appears close enough to linear, with no significant outliers. This suggests a linear model might be reasonable.

The equation of the regression line given in the R output is

$$\widehat{\text{Protein}}(\text{g}) = -1.9692 + 0.0528 \text{ Calories}$$

The scatter plot of residuals vs fitted values exhibits a generally random pattern, with roughly even spread on both sides of the horizontal axis. It seems reasonable to assume the variance is roughly constant. The normal probability plot of the residuals is not perfect, but seems acceptable, particularly with a sample size over 100. Thus, it is reasonable to assume the normal residuals condition is satisfied.

A data.frame: 6 × 12

	Fast.Food.Restaurant	Item	Type	ServingSize_g	Calories	TotalFat_g	SaturatedFat_g	TransFat_g	Sodium
	<fct>	<fct>	<fct>	<int>	<int>	<dbl>	<dbl>	<dbl>	<
1	McDonald's	Hamburger	Burger	98	240	8	3	0.0	
2	McDonald's	Cheeseburger	Burger	113	290	11	5	0.5	
3	McDonald's	Big Mac	Burger	211	530	27	10	1.0	
4	McDonald's	Quarter Pounder with Cheese	Burger	202	520	26	12	1.5	
5	McDonald's	Bacon Clubhouse Burger	Burger	270	720	40	15	1.5	
6	McDonald's	Double Quarter Pounder with Cheese	Burger	283	750	43	19	2.5	

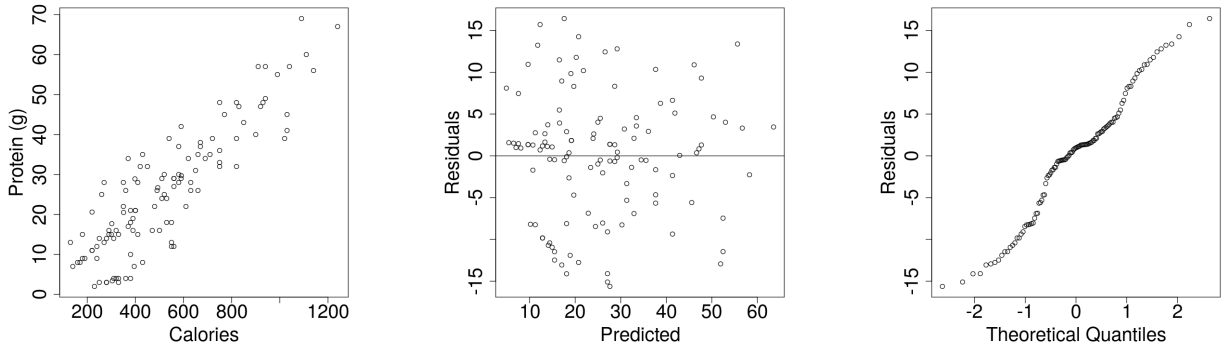
'Fast.Food.Restaurant' · 'Item' · 'Type' · 'ServingSize_g' · 'Calories' · 'TotalFat_g' · 'SaturatedFat_g' · 'TransFat_g' · 'Sodium_mg' · 'Carbs_g' · 'Sugars_g' · 'Protein_g'

```
Call:
lm(formula = Protein_g ~ Calories, data = fooddat)

Residuals:
     Min       1Q   Median       3Q      Max
-15.6199  -5.1625   0.9584   3.8704  16.4193

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.969248   1.561978  -1.261     0.21
Calories     0.052838   0.002677  19.737   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.427 on 116 degrees of freedom
Multiple R-squared:  0.7706,    Adjusted R-squared:  0.7686
F-statistic: 389.6 on 1 and 116 DF,  p-value: < 2.2e-16
```



Shown above is the output from the R code. For clarity, the output produced by `head(fooddat)` has been truncated to show only the first 8 variables. The names of all the variables can be seen in the line below the data table.

To check for independent observations, we would ideally want more information about how the data were collected. But, in the present situation the best we have available is the claim (in the opening sentence of the problem) that the sample is "representative" of fast food items. It is common practice to consider a representative sample as a reasonable alternative

to one that is independent.

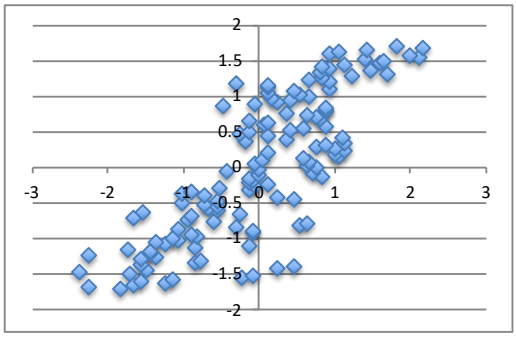
To summarize, it appears that all the necessary conditions are satisfied. Thus, it seems appropriate to use linear regression for predicting protein from calories based on these data. The resulting model is: $\widehat{\text{Protein}}(\text{g}) = -1.9692 + 0.0528$ Calories

The slope indicates that for each additional calorie in a food item, the predicted protein content increases by about 0.05 grams, on average. The regression output also shows that $R^2 = 0.7706$, implying that about 77% of the of the variance in protein content is explained by the variance in calories.

Exercises

1. An ecologist has gathered data on the trunk diameter and age of a species of trees, and found there is a positive linear association between them, with correlation $r = 0.59$. The data satisfies all the conditions for linear regression, and the line of best fit for predicting the age (in years) from the trunk diameter (in inches) has slope of 1.18 with y-intercept of 9.95.
   (a) Identify the explanatory and response variables.
   (b) Using the given information, write the equation of the regression line. Be sure to clarify the meaning and units of your variables.
   (c) Interpret the meaning of the slope in the context of this application (with correct units).
   (d) Find the value of R^2 and interpret its meaning.

2. In this exercise we consider the special case of data sets in which the regression line passes through the origin (see example in the scatter plot). The goal is to find a general expression for the line of best fit, by minimizing the sum of the square of the errors. Accordingly, let the equation of the line be: $\hat{y} = mx$. Assume there are n pairs of data points, denoted by (x_i, y_i). We want to find an expression for m in terms of the (x_i, y_i) values. Here are the steps:



   (a) Write a general expression for e_i, the i^{th} residual. The only unknown it should contain is m. (Note that x_i, y_i are considered known.)
   (b) Compute the function: $f = \sum(e_i)^2$
   This function is the sum of the square of the errors.
   (c) Find the value of m that minimizes f using calculus techniques.
   (d) Write $\hat{y} = mx$, using the above m. This is the required regression line.

Answers: (a) $e_i = y_i - mx_i$. (b) $f = \sum_{i=1}^{n}(y_i - mx_i)^2$.

(c) and (d) $\hat{y} = mx$, with $m = \dfrac{\sum_{i=1}^{n} x_i\, y_i}{\sum_{i=1}^{n} x_i^2}$

3. In regression analysis, it is sometimes useful to re-scale the variables (e.g., by taking logs, or applying other transformations). One common rescaling strategy is to convert all values into z-scores, also known as standardizing. How does standardizing a pair of variables change the corresponding scatterplot? Let us find out.

   (a) Consider the dataset of three (x_i, y_i) values shown in the table. Transform the x-variable into z-scores [i.e., $z_{x_i} = (x_i - \bar{x})/s_x$]. Do the same for the y-variable.

   (b) Compute the mean and standard deviation of each of the transformed variables (z_x and z_y).

   (c) Summarize key features of the z_y vs z_x scatterplot.

x_i	y_i
0	12
1	30
2	28

Answers: (a) $z_x = (-1, 0, 1)^T$, $z_y = (-1.1488, 0.6757, 0.4730)^T$.
(b) The mean = 0 and SD = 1 for all standardized variables.
(c) Since both mean values are 0, the plot will always be centered at the origin. It will have similar shape, strength and direction as the unscaled plot. But the standardized plot will always contain both negative and positive values, even if the original data is all positive. The scatterplot shown in the previous question is a good example of how a standardized plot looks.

4. Consider a dataset consisting of two standardized variables z_x, z_y. Show that in the z_x-z_y plane, the line of best fit is $\hat{z}_y = r z_x$, where r = correlation coefficeint. In other words, the slope of the regression line must equal the correlation coefficeint!
   [Hint: Combine the insights from the previous two questions.]

5. Show that the equation of the regression line for any general pair of x, y variables is

$$\hat{y} = mx + b, \quad \text{with} \quad m = r\frac{s_y}{s_x} \quad \text{and} \quad b = \bar{y} - \left(r\frac{s_y}{s_x}\right)\bar{x}$$

   All other symbols $(r, \bar{x}, \bar{y}, s_x, s_y)$ have their usual meaning.
   [Hint: Start with the result from the previous question and unscale the variables.]

6. A research study explored the relationship between working memory and several other variables in a sample of subjects. Here are summary statistics for working memory capacity (WMC) of the subjects, measured in the form of a unitless numerical score, and their age in years

	Mean	SD
Age (years)	55.6	6.3
WMC (no units)	9.5	3.2
Correlation coefficient, $r = -0.59$		

The researchers want to construct a linear model for predicting WMC from age. Assume the data satisfies all the conditions for linear regression.

(a) Identify the explanatory and response variables (with correct units).

(b) Find the equation of the regression line.

(c) Explain the meaning of the slope in the context of this application.

(d) Find the R^2 value for your model and explain its meaning.

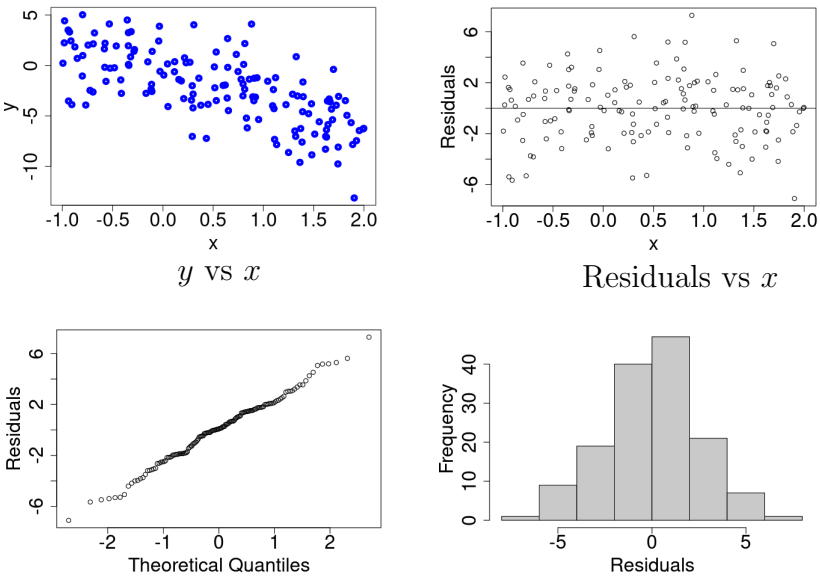
Answers: (a) Explanatory = Age in years. Response = WMC (no units)

(b) Equation of the regression line: $\widehat{\text{WMC}} = 26.16 - 0.2997 \times \text{Age in years}$

(c) Meaning of slope: "For each year that a person ages, the working memory capacity is predicted to decrease by about 0.2997 units, on average."

(d) $R^2 = r^2 = (-0.59)^2 = 0.348$. It means that about 34.8% of the variance in WMC scores is explained by the variance in age.

7. A dataset contains 145 randomized observations of a pair of x, y numerical variables. Determine whether a linear regression model would be appropriate for predicting y from x, based on the diagnostic plots shown in Figure 1.6. Be sure to address all the relevant conditions, and to justify your claims.



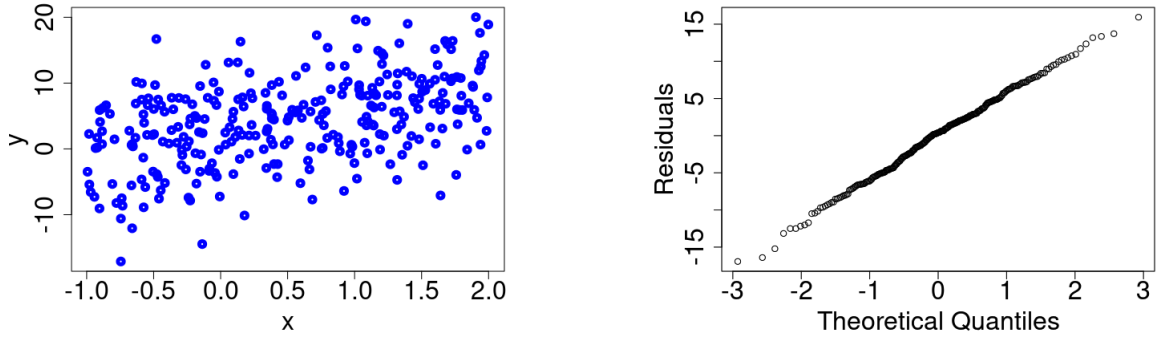
y vs x | Residuals vs x

Residuals normal probability plot | Residuals histogram

Figure 1.6: Diagnostic plots for exercise 7.

8. Another dataset contains 1293 recorded values of a pair of x, y numerical variables. The source of these data is unknown. The only available diagnostic plots from fitting

a linear model to predict y from x are shown in Figure 1.7. Discuss how this dataset does, or does not, satisfy each of the conditions for a valid linear regression model.



y vs x Residuals normal probability plot

Figure 1.7: Diagnostic plots for exercise 8.

Chapter 2

Inferences for Linear Regression

In statistical analysis, the process of developing models and insights using samples is only one step towards understanding the larger reality that we are really interested in studying. When we draw a sample from this larger context (also known as the population) we hope that it captures key features of the underlying population of interest. But, even the best sample will differ from the population, and inference strategies are needed to extend sample-based understanding to population-based understanding.

To exemplify some of the issues of interest, let us revisit the context of Example 2 in the previous chapter. In that example, a sample of 129 high school students was selected to explore whether students' SAT scores were related to their family income. The resulting linear regression model was: $\widehat{\text{SAT score}} = 636.34 + 7.7$ (Family income). However, consider the following real-world issues:

- Do we really want to understand how SAT scores are related to family income just for this sample? Or, do we want to understand that relationship in some larger context?
- Our regression model is the best fit for the **sampled data** – not for the **real world**. How can we find the best regression line for the real world?
- Can we assume the real world relationship is also approximately linear, just from the fact that the sampled relationship is linear?

The goal of this chapter is to address questions such as these, and to introduce inference strategies used in linear regression. We assume the reader has introductory knowledge of basic inference concepts, including strategies such as confidence intervals and significance tests.

Framework and notation

We will assume we are doing simple linear regression between two numerical variables. As discussed in the previous chapter, the general form of the straight line model is

$$\hat{y} = b_0 + b_1 x \tag{2.1}$$

where x is the explanatory variable, $\hat{y}$ is the predicted response, and b_0, b_1 are coefficients that must be found using linear regression on the sampled data.

Let us further assume the sample is drawn from a population in which there is an approximately linear relationship between those same variables. The general notation for the straight line model for the population is usually taken to be

$$\hat{y} = \beta_0 + \beta_1 x \tag{2.2}$$

where β_0, β_1 are the true intercept and slope parameters for the population. In other words, the value of β_0 and β_1 is what we are really seeking, but the best we have available is b_0 and b_1, the sample-based estimates for their value.

Broadly speaking, inferences techniques for regression can be classified into two main categories:

i. Those related to the slope estimate

ii. Those related to the predicted y-values, or response

We will discuss further details in the sections that follow.

Inference methods for the slope

The true slope, β_1, provides a precise, quantitative measure of how the predictor variable affects the response. In many situations, the first question of interest is whether there exists any (linear) relationship between the two variables. Clearly, if there is no relationship then β_1 would be 0. Thus, using the language of inferential statistics, we ask:

Is there sufficient evidence to say that $\beta_1 \neq 0$?

A standard method for answering this question is via a hypothesis test (or, test of significance), which produces a P-value based on which the decision is made. The conceptual framework of this hypothesis test is as follows

- Assume there is no linear relationship in reality: that is, $\beta_1 = 0$.
- But our sample exhibits a linear relationship $\hat{y} = b_0 + b_1 x$, for some $b_1 \neq 0$.
- If β_1 equals 0, what is the probability of a valid sample producing the non-zero b_1 that we see in our sample? This probability is the P-value. If the P-value is sufficiently low, it provides strong evidence to say that β_1 cannot be 0.
- To compute the P-value, we need a probability distribution model for the behavior of b_1 across random samples.

The primary theoretical tool that guides most inference procedures in regression analysis is a probability distribution function that models the behavior of our sampled result. Since the sample is random, it is reasonable to treat each sample-based estimate as a random variable

whose distribution follows some relevant probability model. The rigorous theoretical development of probability models is beyond the scope of the present text. We will primarily rely on statistical software for computations and insights related to probability models. However, we will also strive to retain a big-picture understanding of how and why the inference process works the way it does.

From a big-picture standpoint, the probability distribution function for the slope estimate is the standardized t-distribution, with $n-2$ degrees of freedom. Although we will not compute it by hand, the standard deviation of the slope estimate, which is known as the standard error, is given by

$$SE(b_1) = \sqrt{\frac{1}{n-2} \cdot \frac{\sum(y_i - \hat{y}_i)^2}{\sum(x_i - \bar{x})^2}} = \frac{s_e}{\sqrt{\sum(x_i - \bar{x})^2}} \tag{2.3}$$

where $s_e = \sqrt{\sum(y_i - \hat{y}_i)^2/(n-2)}$ is the standard deviation of the residuals. All symbols have their usual meaning: n = sample size, x_i, y_i = actual value of predictor and response variable for the i^{th} observation, $\hat{y}_i$ = predicted value of response variable, $\bar{x}$ = mean of the predictor variable. Upon closer inspection, it might seem reasonable that $SE(b_1)$ is sometimes interpreted as a measure of the average distance between the regression line and the data points. Clearly, if the data points are far from the line, the numerator of (2.3) would get large, since it captures the variance in the residuals. On the other hand, if the variance in the x-values is large, the denominator of (2.3) gets large, which decreases the value of $SE(b_1)$. Another useful observation from (2.3) is that the units of $SE(b_1)$ are the same as those of the regression slope.

It is important to keep in mind that a valid significance test can only be performed with datasets that satisfy the conditions for linear regression, namely: Linear relationship, normal residuals, constant variance, and independent observations. The procedural steps for carrying out a significance test are summarized below.

Hypothesis test for significance of linear relationship

This test is performed after computing the slope estimate b_1 from the sample. A suitable significance level, say α, is chosen in advance. This is the threshold probability below which a P-value will be considered significant. How to choose α will be discussed later.

1. Null hypothesis: $\beta_1 = 0$. The predictor has no significant effect on the response.
2. Alternate hypothesis: $\beta_1 \neq 0$. The predictor does have an effect on the response.
3. Verify that the data satisfy the required conditions for linear regression.
4. Compute the test statistic using the sample-based estimate

$$ts = \frac{b_1 - 0}{SE(b_1)}, \quad \text{where } SE(b_1) \text{ is provided by software}$$

5. Compute the P-value, which is the probability that $|t| > |ts|$ on the appropriate t-distribution curve. Typically, software output includes the P-value.

6. If P-value $< \alpha$, reject the null hypothesis and conclude the sample provides strong evidence that $\beta_1 \neq 0$. Thus, the predictor does have an effect on the response.
If P-value $> \alpha$, retain the null hypothesis. The sample does not provide evidence of a significant relationship.

The significance level $\alpha \in (0, 1)$ is a probability threshold set by the user. Its choice is typically guided by experience, application context, and tradition. For instance, a very common choice used in many disciplines for virtually every application is $\alpha = 0.05$. This basically implies that if the sample-based estimate (e.g., b_1) occurs with a probability less than 5%, then we are comfortable inferring there is strong evidence of significance. While this may be reasonable for many applications, there is nothing sacred about 5%! Depending on the purpose and context, it is appropriate to use higher or lower α values. Choosing higher values makes it easier to conclude significance, but also increases the risk of that conclusion being wrong (known as a Type I error). Choosing lower values makes it harder to reject the null hypothesis even when it is false (known as a Type II error). Thus, there is no perfect way to choose α. It makes sense to choose very small α values in situations where a Type I error would lead to seriously harmful or dangerous outcomes, and to choose larger values for applications where a Type I error is less mission-critical.

Example 5
An example in the previous chapter looks at the relationship between family income and SAT scores for a random sample of 129 high school seniors in the U.S. The regression line based on that sample was found to be: $\widehat{\text{SAT score}} = 636.34 + 7.7$ (Family income in 1000 \$). We want to know whether family income is a significant predictor of SAT scores in the underlying population of interest (all high school seniors in the U.S.). Carry out a significance test to find out. Assume we are given $SE(b_1) = 0.78$ from software.

Solution
Let β_1 = slope of the true linear relationship between Family income and SAT score in the underlying population.
We will choose $\alpha = 0.1$, since a 10% risk of incorrectly concluding significance seems acceptable in this application context.

* Null hypothesis: $\beta_1 = 0$

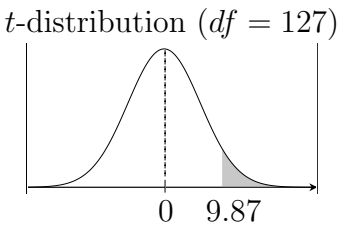
* Alternate hypothesis: $\beta_1 \neq 0$

* In the previous chapter, the data were found to satisfy the required conditions

* The test statistic is

$$ts = \frac{b_1 - 0}{SE(b_1)} = \frac{7.7}{0.78} = 9.87$$

This, essentially, says the sample-based slope is 9.87 standard deviations above 0.



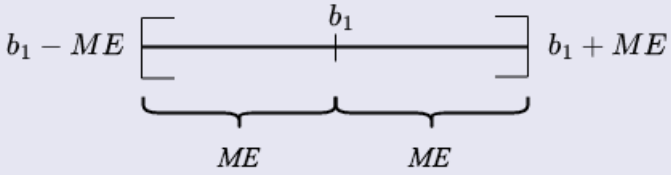
* P-value $= Pr(|t| > 9.87)$ under the t-distribution curve with $df = n - 2 = 127$. The P-value is essentially zero[1] – which is certainly smaller than our α.

* Since the P-value $< \alpha$, we reject the null hypothesis, and conclude the sample provides strong evidence that family income is a significant predictor of SAT scores for the underlying population.

Another standard inference tool for the slope estimate is a confidence interval, which provides a probability based interval that is expected to contain β_1. It is also known as a t-interval for the slope of the regression line.

Confidence interval for the slope of the regression line

Verify that the data satisfy the required conditions for linear regression.
After computing the slope estimate b_1 from the sample, a confidence interval is computed by adding and subtracting an error margin (ME), as shown in this sketch



The margin of error is given by

$$ME = t^*_{n-2} \times SE(b_1)$$

where the value of t^*_{n-2} depends on the sample size (n) and the confidence level we want, and $SE(b_1)$ is the usual standard error of the slope estimate.

The resulting confidence interval is: $CI = b_1 \pm ME$

For instance, suppose we want to compute a 90% confidence interval in Example 5. The

[1]Although one can use a calculator or software to compute this, it is helpful to cultivate a general understanding that the P-value is the area that remains in the tails of the probability distribution. Nine standard deviations is much too far from the center of a z- or t-distribution to retain any measurable area.

calculation step is

$$\begin{aligned} CI &= b_1 \pm t^*_{127} \times SE(b_1) \\ &= 7.7 \pm 1.657 \times 0.78 = [6.408, 8.992] \end{aligned}$$

The value of t^*_{127} was obtained using software. But it is worth noting that with a $df > 100$, the t-distribution is very close to the standard normal distribution. Thus, it is acceptable to use z^* values from a normal curve instead of t^*_{127}.

So, what does this interval mean? One standard way to interpret it as follows:

> With 90% confidence, for each \$1000 increase in family income, the average increase in SAT score for the population (i.e., β_1) lies between 6.408 and 8.992.

It follows that β_1 cannot possibly be 0, according to this confidence interval, which is the same inference we had from the hypothesis test.

In general, we expect both methods to lead to the same inference, provided we use consistent yardsticks for decision-making. This implies, the level of confidence chosen for the interval must be consistent with the α value chosen in the significance test. In our example, we had $\alpha = 0.1$ in a hypothesis test with 2-tails [2] for which 90% is the matching confidence level. More generally, an interval with a confidence level of $100(1 - \alpha)$ will match a 2-tailed test with significance level α.

Inference methods for the predicted response

When a regression model is computed using a sample that satisfies all the necessary conditions, we expect the predicted response at each x to follow a normal distribution. This makes it possible to compute a confidence interval based on the predicted response, also known as a prediction interval. But, there is a subtle issue here that is important to note. It will be easier to see this by revisiting Example 5.

Suppose we want to predict the SAT score of a specific student in that population, say Susan, whose family income is \$70,000. Let $(x_k, \hat{y}_k)$ denote Susan's family income and predicted SAT score. Using the regression model obtained earlier

$$\hat{y}_k = 636.34 + 7.7\ (70) \approx 1175$$

Susan's predicted score is $\hat{y}_k = 1175$, from which we can compute a prediction interval for her true score. It has the usual form of a confidence interval

$$CI = \text{sample-based estimate } \pm t^*_{df} \times SE \text{ of estimate}$$

[2] A test is two-tailed when the alternate hypothesis is $\beta_1 \neq 0$, since the P-value calculation uses both tails of the distribution. In some situations, it may be appropriate for the alternate hypothesis to say $\beta_1 > 0$ (or $\beta_1 < 0$). This is a one-tailed test, with a P-value calculation that uses only the right (or the left) tail.

For a prediction interval based on a sample of size n, the required $df = n - 2$, and the standard error of the estimate is

$$SE_k = s_e \sqrt{1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \tag{2.4}$$

As before, s_e = SD of the residuals, and $\bar{x}$ = mean of the x-variable (family income, in our example). A more detailed understanding of this formula is beyond the scope of the present work. It is sufficient to know that this SE_k approximates the standard deviation of the relevant probability distribution curve, and that we generally expect regression software to provide the numerical values needed to compute it.

In our example, software gives us $s_e = 83.6$. Since all the other quantities are known

$$SE_k = 83.6 \sqrt{1 + \frac{1}{129} + \frac{(70 - \bar{x})^2}{s_x^2 (n-1)}}$$

where we have used the fact that $\sum (x_i - \bar{x})^2 = s_x^2 (n-1)$. We are given $\bar{x} = 75.8$ and $s_x = 9.2$, using which we get $SE_k = 84.1$. The 90% prediction interval for Susan's score is

$$CI_k = 1175 \pm 1.657 \times 84.1 \;=\; [1035.6, 1314.4]$$

Thus, with 90% confidence we predict Susan's SAT score will lie between 1035.6 and 1314.4.

Notice there is a spread of nearly 280 points between the extremes of Susan's predicted score. That is quite large, and makes the prediction less useful from a practical standpoint. Trying to predict a specific individual's score is subject to variability from many sources. This results in a higher standard deviation of the probability distribution curve, which leads to wider confidence intervals. As an alternative strategy, instead of Susan's own SAT score, suppose we were interested in predicting the <u>mean</u> SAT score of all individuals whose family income is \$70,000. A confidence interval for predicting the mean score would have a similar form, but with a smaller SE

$$CI_\mu = \text{sample-based estimate } \pm t^*_{df} \times SE_\mu$$

where

$$SE_\mu = s_e \sqrt{\frac{1}{n} + \frac{(x_s - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \tag{2.5}$$

Plugging in the numerical values from our example

$$SE_\mu = 83.6 \sqrt{\frac{1}{129} + \frac{(70 - 75.8)^2}{9.2^2 (128)}} = 8.7$$

The resulting 90% confidence interval is

$$CI_\mu = 1175 \pm 1.657 \times 8.7 \;=\; [1160.6, 1189.4]$$

With 90% confidence we predict the mean SAT score of all individuals with a family income of \$70,000 will lie between 1160.6 and 1189.4. The width of this interval is about 30 points, in contrast to the 280 points for the prediction interval of an individual's score.

To summarize, there are two types of confidence intervals for the predicted response at a given x-value

1. Prediction interval for an individual response.
2. Confidence interval for the mean response.

The theory, computation and interpretation of these intervals follows the standard strategy used in most confidence intervals. The key difference is in the standard error models (i.e., probability distribution functions), given by equations (2.4) and (2.5). Notice, the equations only differ by an additive constant of 1 inside the square root. But this guarantees that $SE_k > SE_\mu$, implying that a prediction interval for an individual response will always be wider than a confidence interval for the mean response.

We conclude this chapter with a few examples illustrating the use of R software to carry out the needed computations. Note that currently one of the easiest ways to do quick R computations and graphs online (with no login) is through the Sage Cell Server at
`https://sagecell.sagemath.org/`

Example 6
A common task that arises as part of many inference procedures is the need to lookup t-tables for computing probabilities (e.g., P-values), or for doing reverse lookups (e.g., to find critical t^* values). In this example we demonstrate how to use R for t-table lookups in both directions.

a. Find the t^* value for a 90% confidence interval with $df = 20$.
Solution (R code, with output):
```
qt(0.95, df=20)
> 1.724718
```
The first input argument to `qt()` is the cumulative probability, or percentile position. Note that the upper bound on a 90% confidence interval is the 95^{th} percentile position.

b. Find the t^* value for a 98% confidence interval with $df = 73$.
Solution (R code, with output):
```
qt(0.99, df=73)
> 2.378522
```
The upper bound on a 98% confidence interval is the 99^{th} percentile.

c. Find the P-value for a 1-tailed hypothesis test with $df = 20$ and $t = 2.46$.
Solution:
This is essentially asking for the probability $P(t > 2.46)$. The R code, with output is
```
pt(2.46, df=20, lower.tail=FALSE)
```
OR

```
pt(-2.46, df=20)
> 0.0115607
```

The default lookup is always in the lower tail. Since the curve is symmetric around $t = 0$, it follows that $P(t > 2.46) = P(t < -2.46)$.

d. Same as previous Q, except the hypothesis test is 2-tailed.
Solution:
Here we want the probability $P(|t| > 2.46)$. Again, we exploit symmetry and simply double the 1-tailed value. The R code, with output is

```
2*pt(-2.46, df=20)
> 0.0231214
```

e. Find the t-score that cuts off the highest 2.5% of the area when $df = 17$.
Solution (R code, with output):

```
qt(0.975, df=17)   [OR   -qt(0.025, df=17)]
> 2.109816
```

The highest 2.5% is the same as the lowest 97.5%, or the 0.975 percentile position.

Example 7
An example in the previous chapter looked at nutrition data in a sample of 118 fast food items. A linear regression analysis was carried out to predict protein content from the amount of calories, and the resulting model was found to be

$$\widehat{\text{Protein}}(\text{g}) = -1.9692 + 0.0528 \text{ Calories}$$

The following questions assume we have access to the raw data, and can use it to compute all the needed standard errors and intervals.

a. Carry out a hypothesis test to determine whether calories are a statistically significant predictor of protein content.

b. McDonald's has introduced a new Veggie Burger with serving size 200 grams, containing 510 calories. Compute and interpret a 95% prediction interval for the amount of protein in it.

c. Compute and interpret a 95% confidence interval for the mean amount of protein in a fast food item containing 510 calories. Explain how and why this interval differs from the prediction interval in the previous part.

Solution:

a. Let β_1 denote the slope of the true linear relationship between calories and protein. We will choose a significance level of $\alpha = 0.1$.
Null hypothesis: $\beta_1 = 0$
Alternate hypothesis: $\beta_1 \neq 0$
In the previous chapter, the data were found to satisfy the required conditions for

linear regression. Thus, we may proceed with the hypothesis test.

The regression summary contained in the software output (see copy below) shows all the relevant computational results

- standard error for the slope, $SE(b_1) = 0.002677$
- test statistic $= (b_1 - 0)/SE(b_1) = 19.737$
- P-value $< 2 \times 10^{-16}$

```
Call:
lm(formula = Protein_g ~ Calories, data = fooddat)

Residuals:
     Min       1Q   Median       3Q      Max
-15.6199  -5.1625   0.9584   3.8704  16.4193

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.969248   1.561978  -1.261     0.21
Calories     0.052838   0.002677  19.737   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.427 on 116 degrees of freedom
Multiple R-squared:  0.7706,    Adjusted R-squared:  0.7686
F-statistic: 389.6 on 1 and 116 DF,  p-value: < 2.2e-16
```

Since the P-value is well below α, we reject the null hypothesis, and conclude the data provides strong evidence that the calorie content in fast food items is a significant predictor of the protein content.

b. McDonald's new Veggie Burger contains 510 calories. It is possible to compute a prediction interval for protein manually, after computing the standard error using equation (2.4), as we did for the SAT scores example. But, we will demonstrate the use of R in this example. Here is the R code (note that the first 2 lines are not needed if your code has already executed them for the previous exercise)

```
fooddat = read.csv(file="https://cs.earlham.edu/~pardhan/
   sage_and_r/fastfoods.csv", header=TRUE, sep=",")
foodout = lm( Protein_g ~ Calories, data=fooddat)

## The following lines compute the prediction interval
veggieburg = data.frame(Calories=510)
predict(foodout, newdata=veggieburg, interval="prediction",
    level=0.95)
```

Here is the output:

A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	24.97803	10.20601	39.75005

Thus, the 95% prediction interval is: $[10.21, 39.75]$ grams of protein.
Interpretation: With 95% confidence, we predict that McDonald's new 510 calorie Veggie Burger will contain between 10.21 and 39.75 grams of protein.

c. The R code for a confidence interval will be identical to that for a prediction interval, except for a slight change in the last line:

```
predict(foodout, newdata=veggieburg, interval="confidence",
    level=0.95)
```

The resulting output is:

A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	24.97803	23.62169	26.33438

Interpretation: With 95% confidence, we predict the mean protein content in fast food items containing 510 calories lies between 23.62 and 26.33 grams.

This interval differs from the prediction interval in that it provides a much tighter set of bounds for the minimum and maximum values. The width of the interval is barely 3 grams, as it provides an estimate of the mean protein content in all fast food items containing 510 calories. On the other hand, the prediction interval is an estimate of the protein content in a specific food item – McDonald's 510 calorie Veggie Burger. This interval is nearly 30 grams wide, as the variability of protein in a single specific item is much higher than the variability in the mean protein across several items.

Exercises

1. A travel agency wants to study the association (if any!) between domestic airline fares in the United States, and the distance traveled. They collect data on fare and distance for a random sample of 34 flights, and find the scatterplot shows an approximately linear association, with a correlation of $r = 0.53$. Other summary statistics are given in the following table

Fare	mean= \$148, standard deviation= \$48.03
Distance	mean= 1107 miles, standard deviation= 985.4 miles

   (a) Construct a linear regression model to predict fare from distance.

   (b) Compute R^2 and explain what it means in this application context.

   (c) Carry out a hypothesis test to determine whether there is statistically significant evidence of a linear relationship between these variables in the underlying population. Pick any reasonable α. The standard error for the slope estimate is 0.09. Assume the data satisfy all the conditions for linear regression.

(d) Compute a 95% confidence interval based on the slope estimate, and interpret its meaning.

2. An ecologist has gathered data on the trunk diameter and age of a species of trees, and found there is a positive linear association between them. The data satisfies all the conditions for linear regression, and the line of best fit for predicting the age (in years) from the trunk diameter (in inches) has slope of 1.18 with y-intercept of 9.95.

   (a) Write the equation of the regression line. Be sure to identify the variables in your equation, together with their units.

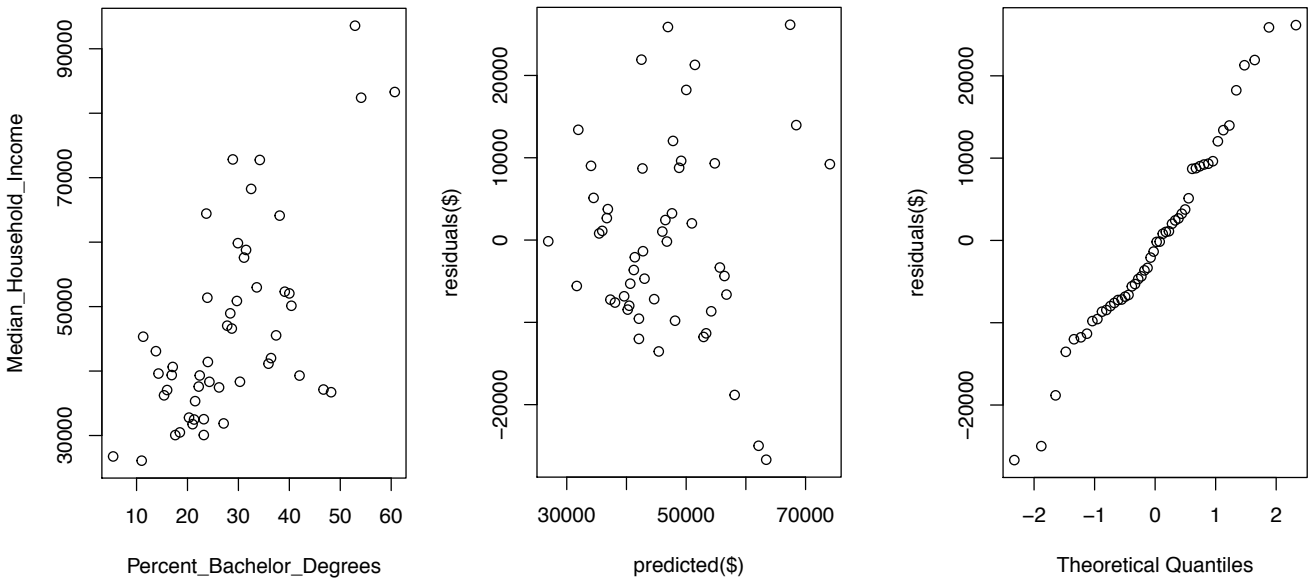
   (b) A regression model for the sampled data is nice! But what relationship is the ecologist, most likely, really interested in studying?

   (c) Assume that relationship is linear, and write its general form using standard notation.

   (d) Write the hypotheses for testing whether there is a statistically significant linear relationship between the variables of interest.

   (e) Suppose the sample size is 37, and the standard error of the slope is 0.24, find the P-value and infer an appropriate conclusion.

3. In this exercise we examine the median income and education level (percent of population with at least Bachelor's degree) for 50 cities in the United States. Shown below are some graphs and summary statistics for constructing a linear regression model to predict median household income (in dollars) from education level. The correlation is $r = 0.65$.



Income	mean= \$46,520, standard deviation= \$15,500
Education	mean= 28.4 %, standard deviation= 12 %

(a) Find the regression model to predict income from education level. Show all steps.

(b) Carefully check the conditions and discuss the appropriateness of your model.

(c) An economist who has been studying similar questions in other regions claims the true slope for this relationship is \$1020 per %. Carry out a hypothesis test to determine whether your result is consistent with this claim. From software, the standard error for the slope estimate is 144.4.