Chapter 15: Describing Relationships: Regression, Prediction and Causation

Overview: In Chapters 14, we began to look into how to describe the relationship between two variables. We used scatterplots to investigate the form and direction of the relationship plotting the explanatory/independent variable on the x-axis and the response/dependent variable on the y-axis. We saw that, when a relationship is linear, we can compute the strength and direction of the linear relationship using correlation r.

In this chapter, we will learn a technique called regression which is used to compute an equation of the form that you believe the relationship has (thus, if you think the relationship is linear, regression is used to compute the equation of the line that best fits the data). We will also discuss how to use the regression model to predict other response values and we'll see why strong correlation does not always imply causation.

Motivating Example: For this chapter, we will work with a dataset containing real estate information. There are many variables in the dataset including, price per unit area, transaction date, house age, distance to nearest transit station, number of convenience stores in the area, latitude, and longitude.

Data Set: Download and save the dataset called RealEstate.csv from Brightspace and load it into R.

Question: What is the response variable in the data set and what are the possible explanatory variables?

Answer: Response variable (y): house price per unit area
Possible explanatory variables: Every other variables in dataset (6 of them)

Save Variables as Vectors: We are going to be working with each of the columns so it is worth while to save them all as shortened variables. We will label our response variable y, and each of the possible explanatory variables as $x1$, $x2$, ...

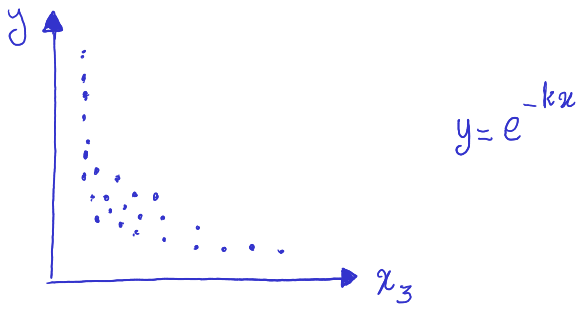
```
y= realEstate$y.house.price.of.unit.area
x1= realEstate$x1.transaction.date
.
.
.
```

Regression with 1 Explanatory Variable:

Suppose we wanted to start by investigating the relationship between the price per unit area (y) and the distance to the nearest transit stations (x3).

Step 1: Make a scatterplot to visualize the relationship and guess the form of the relationship.

plot (x3, y)



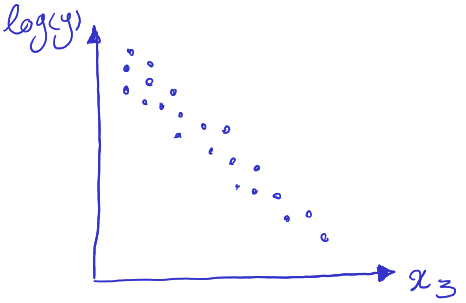
Notice: There appears to be some sort of negative exponential form to the data (when the distance to the nearest transit station is very small, the price per unit area seems to increase significantly and when the distance to the nearest transit station is very large, the price per unit area seems to drop significantly.

Step 2: If the form seems exponential, perform a log-transformation to make the form of the relationship linear.

Recall: If $y = e^{-kx}$, then $\ln(y) = \ln(e^{-kx}) = -kx(\ln e) = -kx$

Plot the log-transformed data and see if this makes the relationship seem more linear.

plot (x3, log(y)) in R log () is same as ln()



Notice: This appears to make the relationship appear more linear, which provides greater support that the original (un-transformed) relationship was negative exponential.

Step 3: We want to try and determine the equation of the line that best represents this relationship.

Definition: A **regression line** is a striaght line that discribes how a response variable y changes as an explnatory value x changes (discribes the linear relationship between x and y)

We often use regression lines to predict the value of y for a given value of x

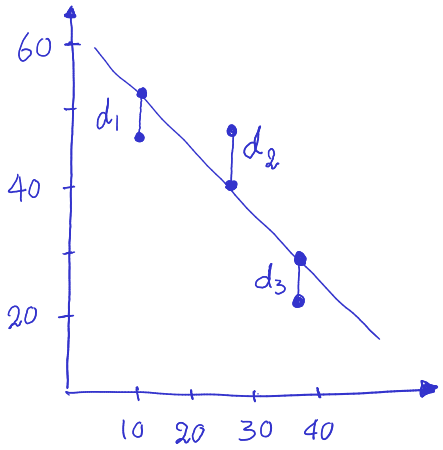
Question: How do we determine this line?

Answer: Using the **method of least squares**.

Definition: The **least squares regression line** is the line that the sum of the squared vertical distance of the data points to the line as small as possible.

Illustrating the Method of Least Squares: Suppose we wanted to find the least squares regression line for the following three pairs of (x, y) data:

$$(20, 60), \quad (35, 40), \quad (40, 20)$$



Minimize: $d_1^2 + d_2^2 + d_3^2$

Determining the Least Squares Regression Line in R: Luckily, we have a command in R that can instantly compute the least-squares regression line: lm()

Example: Determine the equation of the regression line describing the relationship between the log of price per unit area ($\log(y)$) and distance to the nearest transit station ($x3$).

$\log(y) = \beta_0 + \beta_1(x_3)$ (intercept, slope)

x3model= lm(logy ~x3) (response, explanatory variable)

log(y)= 3.820-0.000234 (x3)

Quantities to Describe the Regression Line and the Strength of the Relationship:

We saw in Chapter 14 that we can use correlation r to describe the strength and direction of a linear relationship. There are a few other values that we want to pay attention to when we find a regression line:

- The **square of the correlation** r^2 represents the proportion (or percentage) of the variation in the values of y that is explained by the variation in the values of x (dependent variable; independent variable)

$0 \leq r^2 \leq 1$

For example: It the square of the correlation is 0.64 then if y changes from 50 to 60 (a cahneg of 10 units), then we expect 64% (6.4 units of 10 units) of this change to be explained by x

- The **t-Test on Significance of the coefficients** tells us
  if each term in our regression line is significant or not

  We will not get into all of the details of how to perform a hypothesis test or how to compute the observed value of the t-test but we will give an extremely brief introduction into how to determine if the terms in your regression line are significant or not.

  There are 2 possible options for each coefficient:

  The default option called the **null hypothesis** is that the term is not significant and thus should not be included in the model.

The alternate option called the **alternative hypothesis** is that the term is significant and should be included in the model

If you use $model1 = lm()$ to create a model and then type in the R command: $summary(model1)$

```
Call:
lm(formula = log(y) ~ x3)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7001 -0.1359  0.0083  0.1533  1.0053

Coefficients:
              Estimate  Std. Error t value            Pr(>|t|)
(Intercept)  3.82027067  0.01677884  227.68 <0.0000000000000002 ***
x3          -0.00023395  0.00001009  -23.18 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2589 on 412 degrees of freedom
Multiple R-squared:  0.566,     Adjusted R-squared:  0.565
F-statistic: 537.3 on 1 and 412 DF,  p-value: < 0.00000000000000022
```

The values in the column $Pr(>|t|)$ are called **p-values**. In order to determine if your coefficient is significant use the following rule:

If p-value $\leq$ 0.05, the term is significant (Altervative hyphothesis is true.)
If p-value $>$ 0.05, the term is not significant and should not be in the model.

So in the model relating log of price per unit area ($\log(y)$) with distance to the nearest transit station ($x3$) we see from the summary() function that:

$r^2 = 0.566$ So, 56.6 % of the variation in log(y) can be explained by the variation in x3.

Both p-values for the intercept and slope of x3 are $<$ 0.05, so both term are significant.

Practice Question: Consider now the relationship between price per unit area (y) and latitude ($x5$).

(a) Create two scatterplots, one comparing y and $x5$ and one comparing $\log(y)$ and $x5$. Which one appears to have the stronger linear relationship?

(A) y and $x5$

Plot (x5, y)

(B) $\log(y)$ and $x5$

Plot (x5, log(y))

(b) Determine the linear regression line for your answer to part (a). What is the model?

(A) $y = 19.54 - 484.42(x5)$

(B) $\log(y) = 19.54 - 484.42(x5)$

(C) $y = -484.42 + 19.54(x5)$

(D) $\log(y) = -484.42 + 19.54(x5)$

intercept slope

x5model= lm (log(y) ~ 5)

(c) Are the coefficients all significant?

(A) Yes.

(B) No.

(C) Only the intercept.

(D) Only $x5$.

Summary (X5model)

(d) What percentage of the variation in $\log(y)$ is explained by the variation in $x5$?

(A) 61.7%

(B) 38.19%

(C) 38.04%

log log log

(e) If $\log(y)$ changed from \$3.55 to \$4.00, what dollar amount of that change can we attribute to a change in latitude?

(A) Approx. \$0.20

(B) Approx. \$0.28

(C) Approx. \$0.45

* Change in dependent variable = 4-3.55 = 0.45
Change attributed by change in x5 = 0.45x0.3819≈0.18

- 0.45 ← total
0.18 ← explained by x_5
← 0.27

The remaining 0.27 are explained by other factors that we did not consider in the model.

Using the Regression Line for Prediction:

Once we fit a model to the data, we end up with an equation which we can then use to predict values of the response variable y given values of the explanatory variable x.

Example: We found a model that relates the log of the price per unit area to the latitude position:

$$\log(y) = -484.42 + 19.54(x5)$$

Suppose we wanted to predict the price per unit area when the latitude value is 24.95. Then we could do this by plugging in $x5 = 24.95$ into our regression line:

$\log(y) = -484.42 + 19.54(24.95) = 3.103$ so $y = e^{3.103} = 22.26$ price per unit area

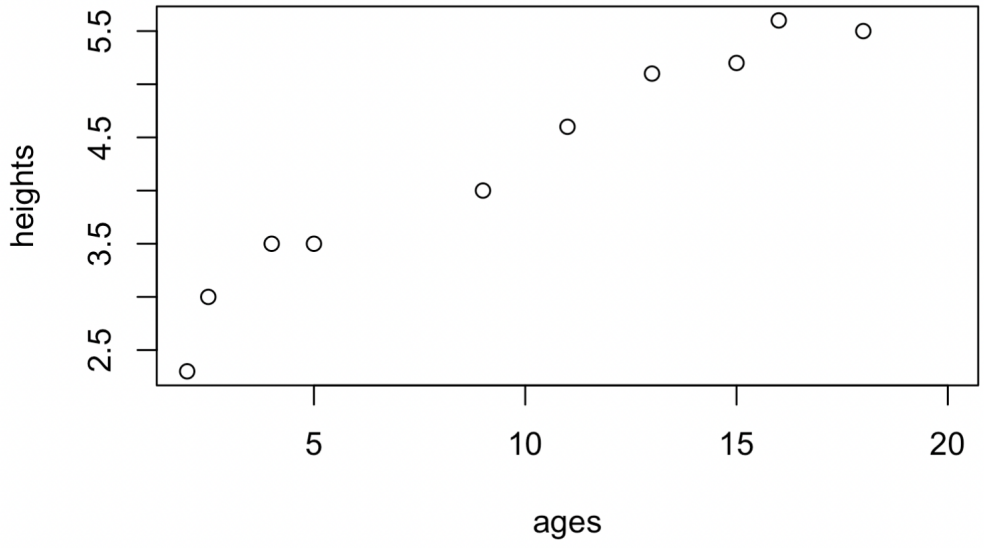
Caution: The regression line was fitted using the data. We do not know if the relationship that was present in the data set will also be present for values outside the data set.

Example: Suppose you wanted to determine the relationship between age and height (in ft). You take a sample of 10 people between the ages 2 and 18 and you get the following results:

$$\text{ages} = (2, 2.5, 4, 5, 9, 11, 13, 15, 16, 18)$$

$$\text{heights} = (2.3, 3, 3.5, 3.5, 4, 4.6, 5.1, 5.2, 5.6, 5.5)$$

Plotting the values we get:



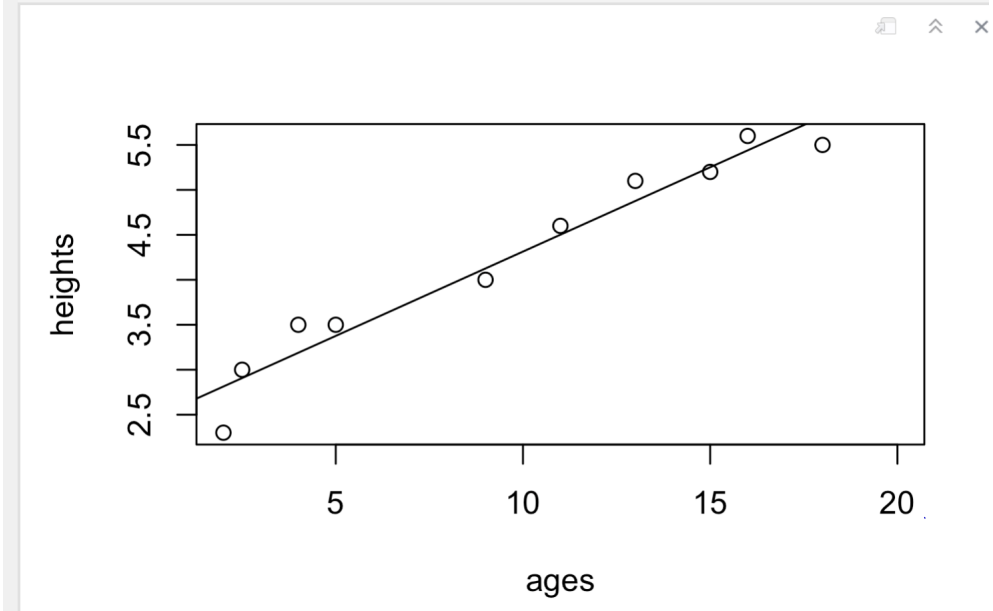
If we use the $lm()$ command we get the following regression line:

```
Call:
lm(formula = heights ~ ages)

Coefficients:
(Intercept)         ages
     2.4385       0.1876
```

We can even plot the regression line with the data using the *abline*() command in R:

```
plot(x=ages,y=heights,xlim=c(2,20))
abline(lm(heights ~ ages))
```



We could use this model to predict the height of an individual who is 12 years old:

heights = 2.4385 + 0.1876 (ages)

heights = 2.4385 + 0.1876 (12) = 4.897 ft

But if we try and use this model to predict the height of an individual who is 40 years old:

heights = 2.4385 + 0.1876 (40) = 9.9425 → not realistic

Definition: **Extrapolation** is

Extrapolation is incredibly risky (as we saw in the previous example). This is why models which try to predict future events (such as economic trends) are often wrong and must be used with extreme caution.

Correlation vs Causation:

Another thing that we need to be careful with when exploring the relationship between variables is equating a strong correlation between variables with a causal relationship between the variables.

Example: An observational study is performed where data on average life expectancy and average number of television sets per person in various countries is collected.

It was observed that countries with higher average numbers of television sets per person also had higher life expectancies. The positive linear relationship was strong with a correlation of $r = 0.85$.

Question: Does this mean that owning more televisions increases your life expectancy?

Answer: Probably not, the underling causes of incresed life expectancy likely has more to do with income level (or wealth level) since increased income means increased access to health resources and increased access to luxuries such as TVs.

Possible Explanations for a Relationship between variables: Some explanations for observed association between variables are:

we will use $\longrightarrow$ to represent a cause and effect link
and $\cdots\cdots$ to represent an association between variables

1. **Causation:**

e.g x = # of hours of sleep
y = energy level during the day

x occurs and it directly causes something to happen to y.

even when direct causation is present, it rarely is a complete explanation of a relationship between variables

Regression with Multiple Explanatory Variables: We can also perform linear regression with multiple explanatory variables.

Example: Suppose we wanted to determine. a model illustrating the relationship between price per unit area (y) and both the distance to the nearest transit station ($x3$) and the latitude ($x5$). Again, we use the $lm()$ function in R:

independent variable independent variable

model_x3_x5= lm(y~x3+x5)

Which gives us the model:

y= -6189.467-0.005811(x3)+249.659(x5)

Looking at the $summary()$ of the model, we see that:

1. p-value for each item is less than 0.05. So they are all signifanct.
2. The square of the correlation is 0.4875. So 48.75% of the variation in y is expected by the variation in x3 and x5

Practice Question: Fit a linear regression model using all 6 possible explanatory variables. Look at the summary of the model, which variables (if any) are not significant (and thus should not be included in the model)? Select all that apply.

(A) $x1$ (C) $x3$ (E) $x5$

(B) $x2$ (D) $x4$ (F) $x6$

Now fit a new model, including only the variables that were significant. The model is:

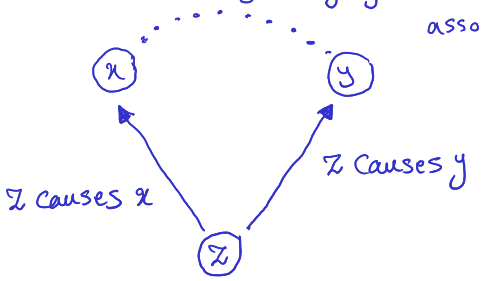
Use the model to predict the price per unit area of a property with transaction date: 2012.667, house age: 20, distance to nearest transit station: 120, number of convenience stores: 3, latitude: 24.97, and longitude: 121.5.

y = -15959.26+ 5.135* 2012.667-0.269*28+-0.004*120+1.136*3+226.882*121.5= 38.5 price per unit area

2. **Common Response:**

e.g x = # of TV sets
y = life expectancy
z = income

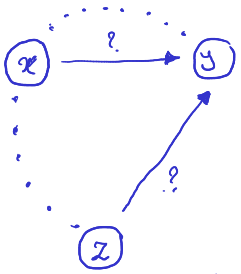
we see both x and y changing at the same time. So, we see an association between them



3. **Confounding:**

e.g x = income
y = life expectancy
z = family medical history

There is an association between x and y



There is an association between x and z

We cannot distinguish the influence of x from the influence of z on y

Question: How do we know that there is a true causal relationship?

Answer: We don't know for sure, however, there are some things which can build a strong case for a causal relationship between variables.

- The association is strong.
- The association is consistent (across multiple studies).
- The alleged cause precedes the alleged effect in time.
- The alleged cause is plausible.
- Experiments holding other possible explanatory variables constant, still show a strong association.

Linear Regression with Non-Linear Models: If we suspect the form of the relationship between our response variable and our explanatory variable is non-linear (for example, it might look quadratic) then we can use linear regression to come up with the quadratic model.

Note: This brings up an important point about what the word **linear** in linear regression is referring to. It is in reference to the coefficients of the explanatory variables, not the variables themselves.

Example: Consider the following two models and determine which one is linear and which one is not

$$y = \beta_0 + \beta_1 x_1 + \frac{\beta_2}{\beta_3} x_2$$

non-linear

We can estimate β_0 and β_1, but we can't estimate β_2 and β_3, seperately.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 (x_1)^2 + \beta_3 (x_2)$$

linear

x_1, x_2

$x_3 = x_1^2$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \beta_3 x_2$$

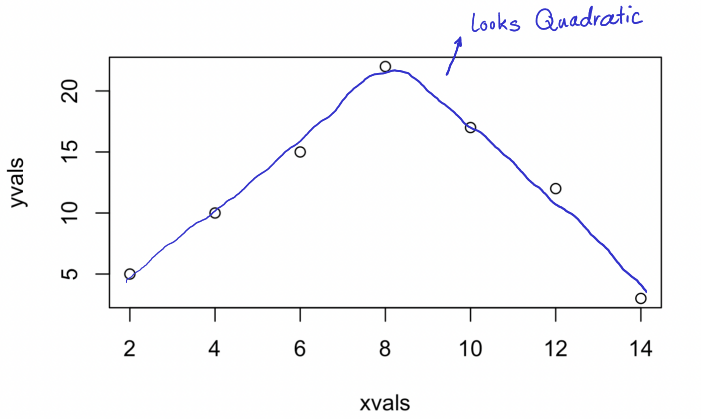
Modelling a Quadratic Relationship:

Example: Suppose you are looking to model the relationship between a response variable y and an explanatory variable x. You have the following observed values of the variables:

```
xvals = c(2,4,6,8,10,12,14)

yvals = c(5,10,15,22,17,12,3)
```

When we plot the values, we see that the relationship appears quadratic:



So we can try and fit the following model to the data:

$$y = \beta_0 + \beta_1 x + x^2$$

Using the following R command:

$x^2 = x \times x$

```
xsq = xvals*xvals

quad_model = lm(yvals ~ xvals + xsq)
```

Then, using $summary(quad\_model)$ we see:

sammary(quad_model)

The p-value for both independent variables are less than 0.05. So, we can keep both variables.

* Should be re-run the model without the intercept?

Usually in practice, we leave intercept alone. If you decide to remove intercept. You need to re-run the following model:

quad_model = lm(yvals ~ 0+xvals+xsq)

And we can use the model to predict other response values. For example, if $x = 11$, then:

$$-8.5714 + 6.8571(11) - 0.4286(11)^2 = 14.9961$$