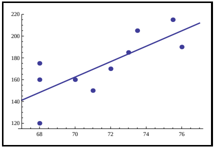
Another R^2 Calculation

Example. Estimating weight from height.

To the right is a list of heights and weights for ten students. We can calculate the line of best fit:

$$(\text{weight}) = 7.07(\text{height}) - 333.$$



ht.	wt.
68	160
70	160
71	150
68	120
68	175
76	190
73.5	205
75.5	215
73	185
72	170

Now find the correlation coefficient: $(\overline{w} = 173)$

$SSE = \sum_{i=1}^{10}(w_i - [(7.07)h_i - 333])^2 \approx 2808$

$SST = \sum_{i=1}^{10}(w_i - 173)^2 = 6910$

So $R^2 = 1 - (2808/6910) = 0.59$, a good correlation.

We can do better by introducing another variable:

Multiple Linear Regression

Add waist measurements to the list:

We wish to calculate a relationship such as:

$$(\text{weight}) = a(\text{height}) + b(\text{waist}) + c.$$

This is no longer a line; it is a best-fit plane.
We can still apply least-squares criterion. Minimize:

$$SSE = \sum_{(h_i, ws_i, wt_i)} \left[wt_i - (a \cdot h_i + b \cdot ws_i + c)\right]^2$$

To find that the best fit plane is (coeff sign)

$$(\text{weight}) = 4.59(\text{height}) + 6.35(\text{waist}) - 368.$$

ht.	wst.	wt.
68	34	160
70	32	160
71	31	150
68	29	120
68	34	175
76	34	190
73.5	38	205
75.5	34	215
73	36	185
72	32	170

Compare the predicted value for the one-variable regression:
$$[\hat{z}_1 = 7.07 \cdot 68 - 333 = 160.02]$$
with the results for two-variable regression
$$[\hat{z}_1 = 4.59 \cdot 68 + 6.35 \cdot 34 - 368 = 147.76]$$

Multiple Linear Regression

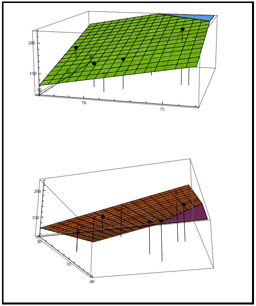
Visually, we can see that we might expect a plane to do a better job fitting the points than the line.

- Now calculate R^2.

Calculate $SSE = \sum_{i=1}^{10}(w_i - f(h_i, ws_i))^2 \approx 955$

SST does not change: (why not?)

$\sum_{i=1}^{10}(w_i - 173)^2 = 6910$



ht.	wst.	wt.
68	34	160
70	32	160
71	31	150
68	29	120
68	34	175
76	34	190
73.5	38	205
75.5	34	215
73	36	185
72	32	170

So $R^2 = 1 - (955/6910) = 0.86$, an excellent correlation.

Notes about the Correlation Coefficient

Example. Cancer and Fluoridation. (pp. 188–189)
Does fluoride in the water cause cancer?

Variables:
$L =$ log of years of fluoridation $A =$ % of population over 65.
$C =$ cancer mortality rate

Use a linear regression to find that
$C = \mathbf{27.1}L + 181$, with an $R^2 = 0.047$.

Compare to a multiple linear regression of
$C = \mathbf{0.566}L + 10.6A + 85.8$, with an $R^2 = 0.493$.

- Be suspicious of a low R^2.
- Signs of coefficients tell positive/negative correlation.
- Cannot determine relative influence of one variable in one model without some gauge on the magnitude of the data.
- Can determine relative influence of one variable in two models.

Notes about the Correlation Coefficient

Example. Time and Distance (pp. 190)
Data collected to predict driving time from home to school.

Variables:
T = driving time S = Last two digits of SSN.
M = miles driven

Use a linear regression to find that
$T = 1.89M + 8.05$, with an $R^2 = 0.867$.

Compare to a multiple linear regression of
$T = 1.7L + 0.0872S + 13.2$, with an $R^2 = 0.883$!

- R^2 increases as the number of variables increase.
- This doesn't mean that the fit is better!

Modeling: Start to Finish

Example. Vehicular Stopping Distance

Background: Back when you took driver's training, you learned a rule for how far behind other cars you are supposed to stay.

- Stay back one car length for every 10 mph of speed.
- Use the two-second rule: stay two seconds behind.

This is an **easy-to-follow** rule; it is a **safe** rule?

State the question:

1. Does the two-second rule fit the 10 mph rule?
2. Does the two-second rule mean we'll stop in time?
3. Determine the total stopping distance of a car as a function of its speed.

Identify factors:

Stopping distance is a function of what?

Breaking down the problem

Describe mathematically and do mathematical manipulations:

Subproblem 1:

Determine reaction distance

Assume speed is constant throughout reaction distance. Then total reaction distance is $d_r = t_r \cdot v$.

Subproblem 2:

Determine stopping distance

Assume brakes applied constantly throughout stopping, producing a constant deceleration.

Brake force is $F = ma$, applied over a breaking distance d_b.

This energy absorbs the kinetic energy of the car, $\frac{1}{2}mv$.

Solve $m \cdot a \cdot d_b = \frac{1}{2}mv^2$ to find that we expect $d_b = C \cdot v^2$.

Total stopping distance is therefore $d_r + d_b$.

Model verification

Model Evaluation:

- Did we answer the question?
- Can we gather data?
- Does it make sense?
- If so, collect data in order to find the constants.

Data is available from US Bureau of Public Roads. (Fig. 2.14)

The data lie perfectly (**!**) on a line. $d_r \approx 1.1v$.

- Examine methodology of data collection.
- Experimenters said $t_r = 3/4$ sec and calculated d_r!
- Perhaps we should design our own trial?

Model verification

- Data for braking distance is a range.
  - Trials ran until had a large enough sample
  - Then middle 85% of the trials given.
- We're modeling d_b as a function of v^2, so transform the x-axis.
- Do we try to fit to low value, avg value, or high value in range?
  - Goal: prevent accidents!

Consider the line in Figure 2.15:

$$d_b = 0.054v^2.$$

Up to 60 mph, line seems like reasonable it.

- *Conclusion:* $d_{tot} = d_r + d_b = 1.1v + 0.054v^2$.
- Check fit by comparing plots of observed stopping distance and model's predicted stopping distance (Fig. 2.16)

Decide model is reasonable at least until 70 mph.

Limitations and assumptions inherent in our model:

When is our model reasonable?

- Drivers going ≤ 70 mph
- Good road conditions
- Driving car, not truck
- Current car manufacturing

Implement the model

- Come up with a good rule of thumb for drivers to follow (Next slide!)
- Publicize it

Maintain the model

- Revise every five years
- In the future, perhaps there will be no accidents!

Vehicular Stopping Distance

What about that two-second rule?

- **Easy to implement.**
- Two-second rule is a linear rule,
- A quadratic rule would make more sense.
- Works up until 40 mph, then quickly invalid! (Fig 2.17)

Come up with a variable rule based on speed.

- It's not reasonable to tell people to stay 2.5 seconds behind at 50 mph and 2.8 seconds behind at 58 mph!
- Determine speed ranges where
  - two seconds is enough ($\leq$ 40 mph)
  - three seconds enough ($\leq$ 60 mph)
  - four seconds enough ($\leq$ 75 mph)
  - And more if non-ideal road conditions.