

SLR & CORRELATION

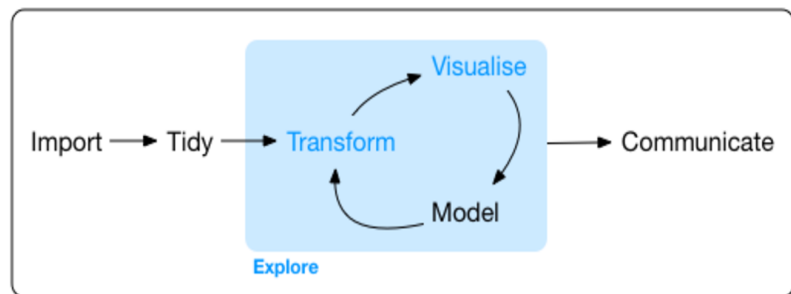
Research Methods in Psychology I & II • Department of Psychology • Colorado State University

BY THE END OF THIS UNIT YOU WILL:

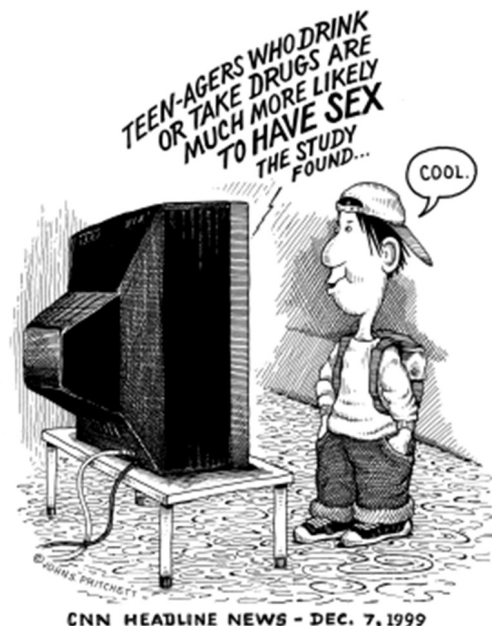
1. Understand the utility of relating two variables with a straight line.
2. Know how to use the least squares criterion to find the best fit line that relates one variable to another.
3. Be able to define this best fit line with an equation.
4. Understand the relationship between correlation and SLR.
5. Know how to determine if SLR parameter estimates are statistically different from zero.
6. Know how to evaluate and present the precision of SLR parameter estimates.
7. Know how to fit a Simple Linear Regression (SLR) in R.
8. Be proficient with SLR plotting procedures in R.

What is Simple Linear Regression (SLR)?

SLR is a statistical method that allows for some outcome variable (y) to be regressed on some predictor variable (x), in a way that allows the user to determine the extent to which x can predict y .



Wickham & Grolemund—R for Data Science

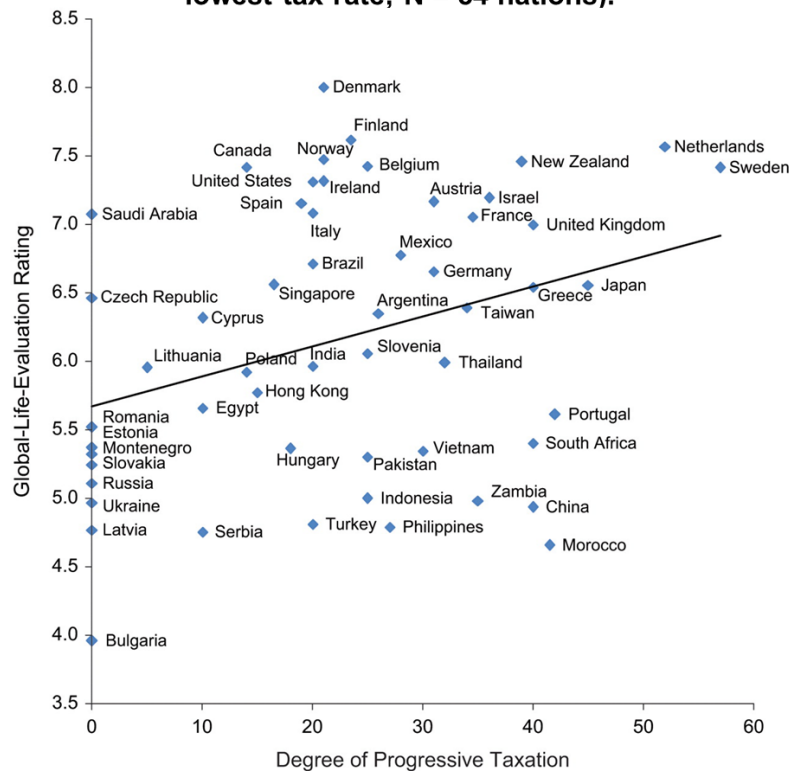


CNN HEADLINE NEWS - DEC. 7, 1999

What is Simple Linear Regression (SLR)?

- In social science, we often want to know whether two or more variables are associated with one another.
- By associated we mean that knowing the value of one variable tells us something about the value of the other variable.
- Simple linear regression helps us to understand the extent to which two variables are associated by estimating the straight line that relates them to one another.
- Let's consider an example.

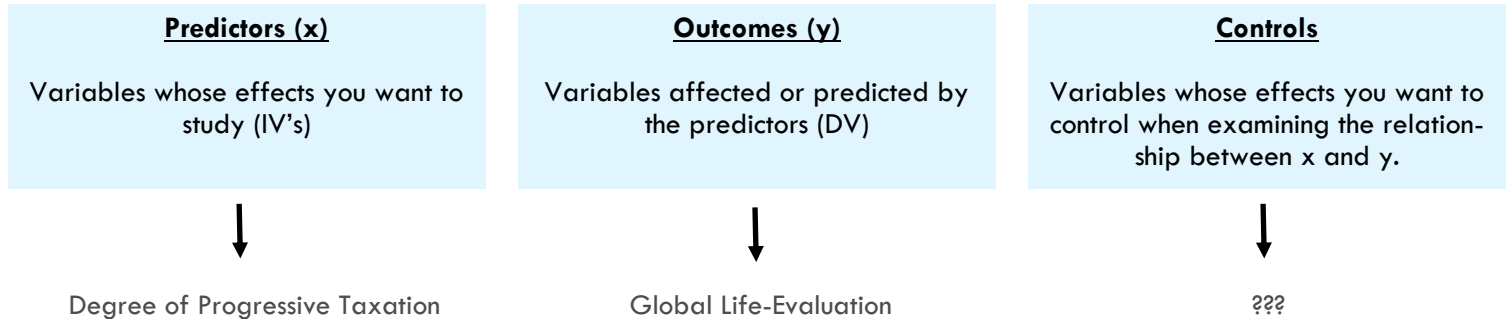
Fig. 1. Scatter plot (with best-fitting regression line) showing mean global-life-evaluation rating as a function of progressive taxation (calculated as the difference between the highest and lowest tax rate; N = 54 nations).



Oishi, S., U. Schimmack, and E. Diener (2012). Progressive Taxation and the Subjective Well-Being of Nations. *Psychological Science*, 26(1), 86-92.

"The influential philosopher John Rawls (1971/1999) argued that a good society redistributes its wealth so that all its citizens have the same opportunity for future success in the form of equal access to public goods, such as quality education and health care. According to Rawls, then, progressive taxation (a higher tax rate for the rich than for the poor) is a more just policy than flat taxation (a constant tax rate across income categories). Using data from the Gallup World Poll, we examined whether progressive taxation is associated with the subjective well-being of nations."

Types of Variables in Linear Regression Models:



Types of Models

Mathematical Models

Mathematical models are **deterministic**. Once the “rule” is known, it can be used to **perfectly** fit the data. That is, we can account for ALL of the variance in the outcome.

Examples:

1. Perimeter of a square = $4 \times (\text{length of a side})$
2. Area of a square = $(\text{length of a side})^2$

Statistical Models

Statistical models are **not deterministic**. They take into account that we usually don't have all important predictors, that we rarely perfectly measure the variables that we do have, and that people (or organizations, schools, animals, etc.) act differently.

Statistical models allow for:

1. Excluded variables
2. Measurement error
3. Individual variation

These are the kinds of models that we will explore in this course.

$$\text{Outcome} = \text{Systematic Component} + \text{Residual}$$



AN EXAMPLE

Successful Weight Loss

A research team is interested in assessing the relationship between caloric deficit and pounds lost over the course of a 1 month weight loss program designed for men age 40 to 50 in CO. 100 obese men were randomly selected to participate. She measured the following variables:

1. **lbslost** = pounds lost after one month on the weight loss program
2. **caldeff** = caloric deficit over the course of the program (expressed in 1000 calories). A caloric deficit is a state in which you are burning more calories than you eat.
3. **selfeff** = self efficacy for weight loss at the start of the program

DATASET: wtloss.csv

The research team wants to assess the relationship between caloric deficit and pounds lost over the course of the weight loss program. Their hypothesis is that men who accumulate a greater caloric deficit will lose more pounds.

Please set up a new notebook, and call it WeightLoss_Notebook. Save it in your MyClassActivities project folder. Begin the notebook with the following chunks. Note that there is one new package, so you will need to install it first:

```
install.packages("olsrr")
```

```
WeightLoss_Notebook.Rmd
5
6 This is a R Notebook to explore the wtloss dataframe
7
8 # Load libraries
9 ```{r}
10
11 library(olsrr)
12 library(descriptr)
13 library(psych)
14 library(modelr)
15 library(tidyverse)
16 library(mosaic)
17
18 ```
19
20 # Import data
21 ```{r}
22
23 wtloss <- read_csv(file = "wtloss.csv")
24
25 ```
26
27 # Subset to keep just variables we need
28 ```{r}
29
30 wtloss <- select(wtloss, lbslost, caldef)
31
32 ```
```

Let's Request Univariate Descriptive Statistics and Basic Plots

```
summary_stats(wtloss$lbslost)
```

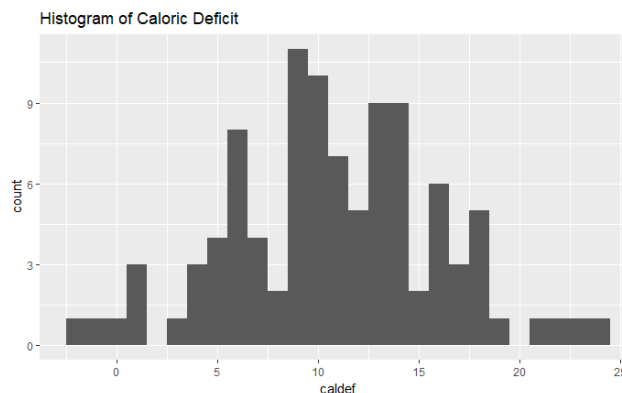
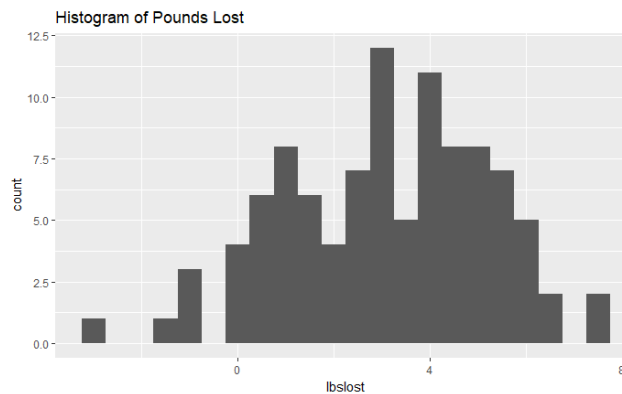
```
summary_stats(wtloss$caldef)
```

```
ggplot(wtloss, aes(x = lbslost)) +  
  geom_histogram(binwidth = .5) +  
  labs(title = "Histogram of Pounds Lost")
```

```
ggplot(wtloss, aes(x = caldef)) +  
  geom_histogram(binwidth = 1) +  
  labs(title = "Histogram of Caloric Deficit")
```

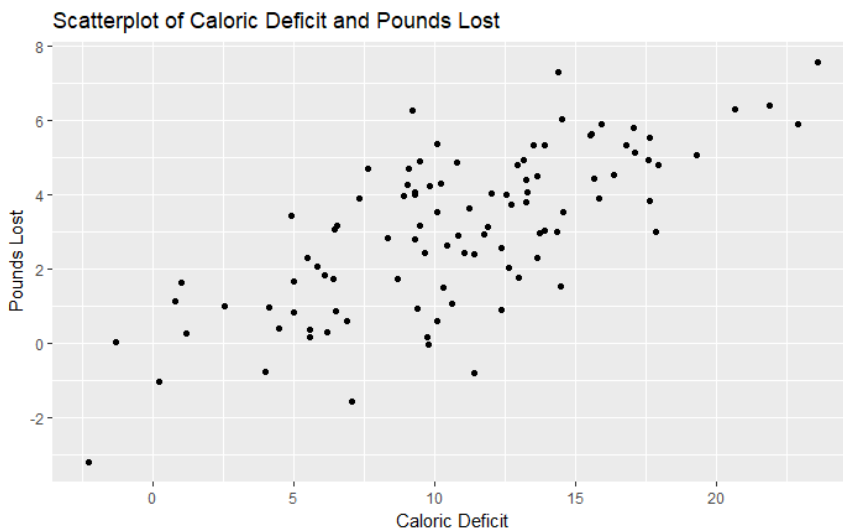
Pounds Lost (lbslost)	N	100.00	Variance	4.40
	Missing	0.00	Std Deviation	2.10
	Mean	3.06	Range	10.76
	Median	3.14	Interquartile Range	3.07
	Mode	-3.19	Uncorrected SS	1374.33
	Trimmed Mean	3.11	Corrected SS	435.83
	Skewness	-0.33	Coeff Variation	68.49
	Kurtosis	-0.22	Std Error Mean	0.21

Caloric Deficit (caldef)	N	100.00	Variance	26.40
	Missing	0.00	Std Deviation	5.14
	Mean	10.80	Range	25.84
	Median	10.69	Interquartile Range	6.64
	Mode	-2.27	Uncorrected SS	14286.03
	Trimmed Mean	10.82	Corrected SS	2613.56
	Skewness	-0.08	Coeff Variation	47.56
	Kurtosis	0.13	Std Error Mean	0.51



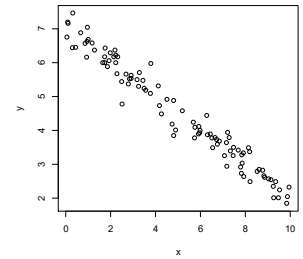
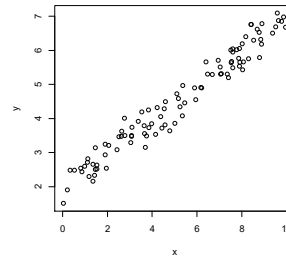
Let's Request a Scatterplot of Pounds Lost & Caloric Deficit

```
ggplot(wtloss, aes(x = caldef, y = lbslost)) +  
  geom_point() +  
  labs(title = "Scatterplot of Caloric Deficit and Pounds Lost", x = "Caloric Deficit", y = "Pounds Lost")
```

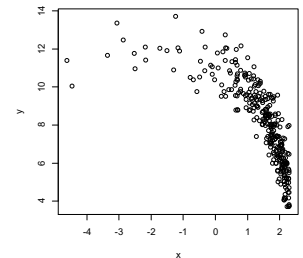
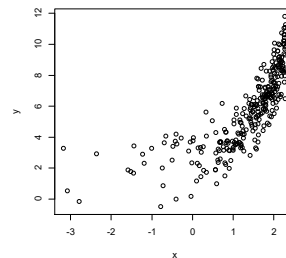


Always plot the data before you estimate a model. Here are some key questions to ask when examining scatter plots.

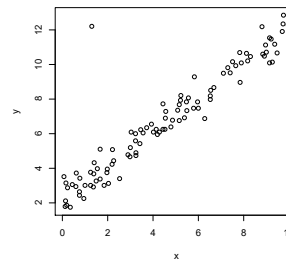
What is the direction of the relationship?



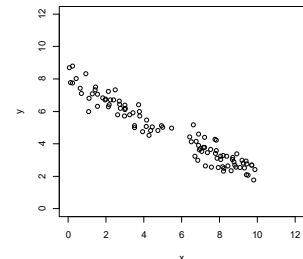
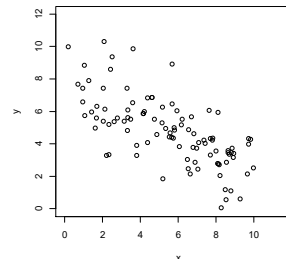
Is the relationship linear?



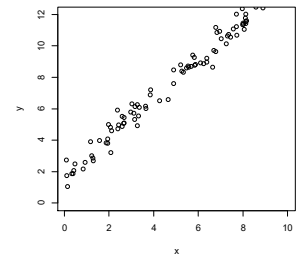
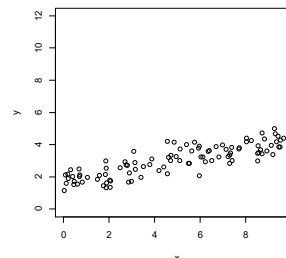
Are there any outliers?



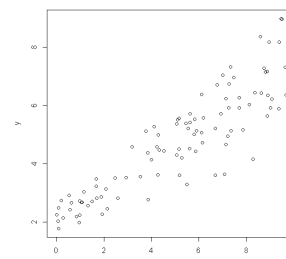
How strong is the relationship?



What is the magnitude of the relationship?



Is the distribution of y at each value of x similar?



What is the direction of the relationship?

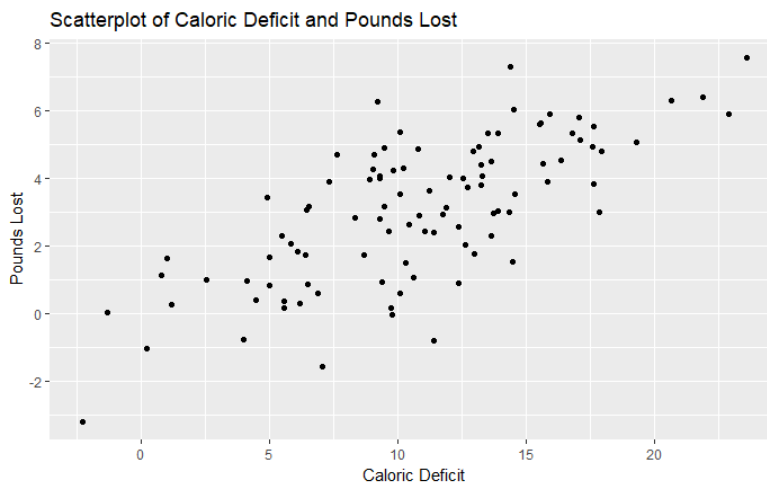
POSITIVE

Is the relationship linear?

YES

Are there any outliers?

NO



How strong is the relationship?

MODERATE TO STRONG

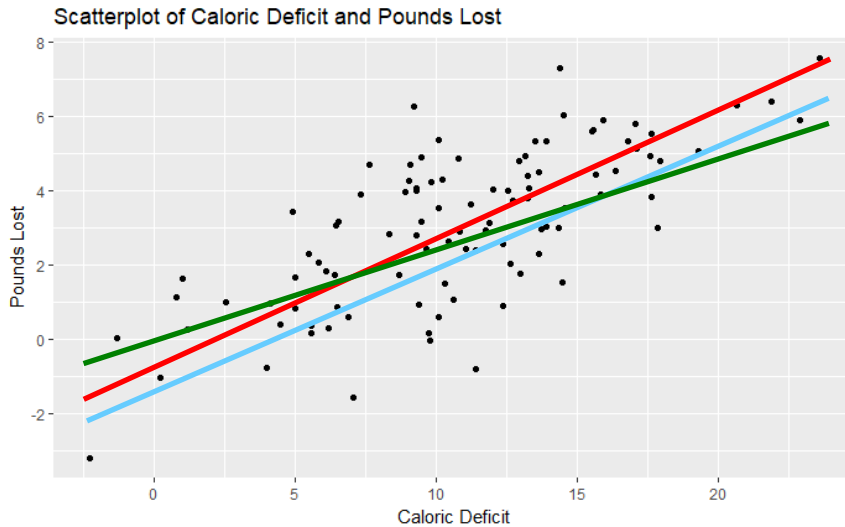
Is the distribution of y at each value of x similar?

YES

What is the magnitude of the relationship?

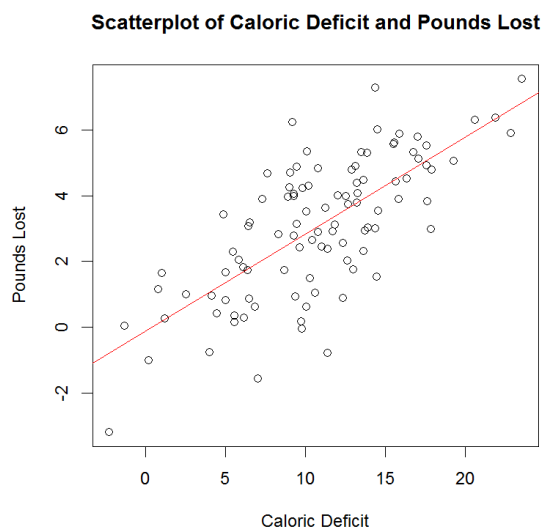
SLOPE $\sim .29$

It looks like there is a positive relationship. That is, the greater caloric deficit that one accumulates, the more weight they lose. However, we need a method for determining the equation of the line that would describe this positive relationship – i.e., the besting fitting line for these data.



It is possible to fit an infinite number of straight lines to the data points in the scatter plot. However, we want to find the “best fitting line.”

The best fitting line is the line that allows x (caloric deficit) to predict y (pounds lost) with the highest amount of accuracy. This best fitting line is defined by an equation consisting of an intercept (the predicted value of y when $x=0$) and a slope (the predicted change in y for each one unit increase in x).



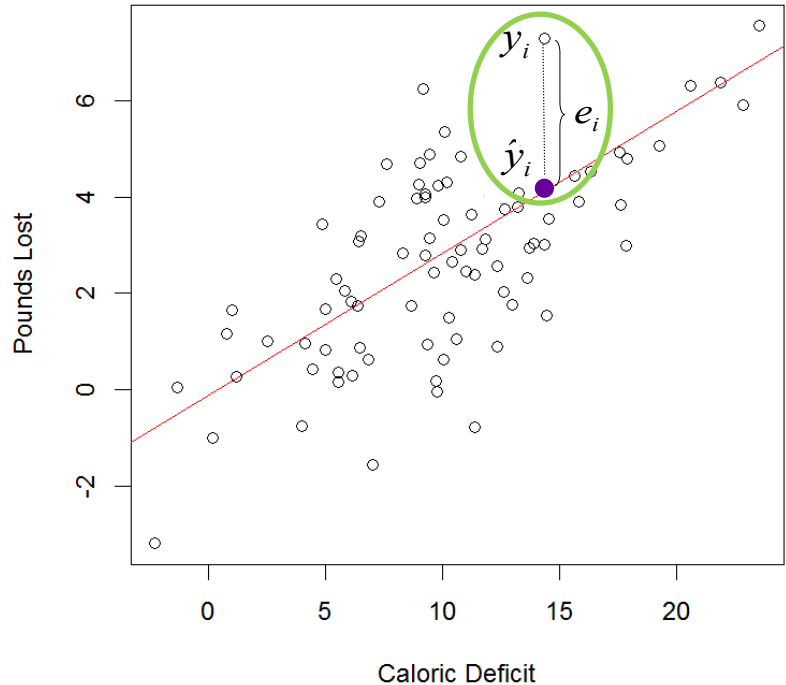
$$y_i = b_0 + b_1 x_i + e_i$$

intercept
slope

We use the least squares criterion to find the best fit line. This process starts by considering each case's residual. The residual is the difference between the observed value of y and the predicted value of y (the value that falls on the regression line — this is called \hat{y}).

Each case (e.g., individual—represented as i) in the dataset has an observed y , a \hat{y} , and a residual.

Scatterplot of Caloric Deficit and Pounds Lost



y_i Observed y (e.g., observed pounds lost)

\hat{y}_i Predicted y (e.g., the value of y that we predict based on the individual's score on x) - the point on the best fit line

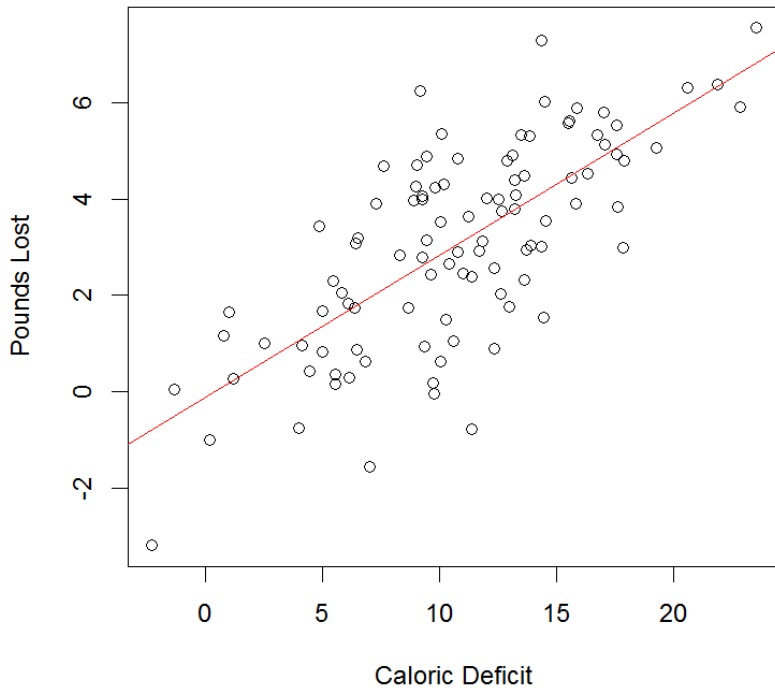
e_i Residual—calculated as observed y minus predicted y

The case under consideration here (i.e., in the green circle) is ID#124. This individual had a caloric deficit of 14,360 and they lost just over 7 pounds. Case #124's predicted weight loss, given their caloric deficit, is about 4 pounds. This line under predicts their observed score. We can calculate their residual by subtracting their predicted score from their observed score—this yields 3.18.

id	lbslost	caldef	\hat{y}_i	e_i	e_i^2
101	4.71	9.07	2.55	2.16	4.65
102	2.78	9.30	2.62	0.16	0.03
103	4.93	17.58	5.06	-0.13	0.02
.					
124	7.29	14.36	4.11	3.18	10.11
.					
.					
.					
199	3.18	6.55	1.81	1.37	1.88
200	5.06	19.31	5.57	-0.51	0.26

The best fitting line is the line that results in the smallest sum of squared residuals (i.e., each case's residual is squared and then all squared residuals are summed).

$$\Sigma = 208.80$$

Scatterplot of Caloric Deficit and Pounds Lost

Any line drawn through the data points, including the best fitting line as determined by the LSC, can be written in equation form.

The residual (e_i) is the difference between an individual's observed and predicted score on y .

$$y_i = b_0 + b_1x_i + e_i$$

The intercept (b_0) is the predicted value of y when $x=0$.

The slope (b_1) is the predicted change in y for a 1 unit increase in x .

id	lbslost	caldef	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	\hat{y}_i	$(\hat{y}_i - \bar{y})^2$	$(y_i - \hat{y})^2$	$(y_i - \bar{y})^2$
101	4.71	9.07	-2.85	2.99	2.55	0.26	4.65	2.71
102	2.78	9.30	0.42	2.26	2.62	0.20	0.03	0.08
103	4.93	17.58	12.67	45.90	5.06	3.99	0.02	3.50
104	6.25	9.19	-5.14	2.60	2.59	0.23	13.40	10.14
105	1.73	8.70	2.79	4.41	2.44	0.38	0.50	1.77
.								
.								
.								
200	5.06	19.31	16.95	72.30	5.57	6.28	0.26	3.97

Mean 3.06 10.80

Σ 770.30 2613.56

In order to determine the best fitting line, we rely on two equations called **the normal equations**.

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{770.30}{2613.56} = .29 \quad \leftarrow \text{Equation for the slope}$$

$$b_0 = \bar{y} - b_1 \bar{x} = 3.06 - .29(10.80) = -.12 \quad \leftarrow \text{Equation for the intercept}$$

$$y_i = -.12 + .29x_i + e_i$$

This value (the residual) accounts for the fact that we do not perfectly predict each individual's score.

Among men who accumulate no caloric deficit (i.e., $x=0$), we predict .12 pounds of weight gain during the course of the program.

For each one unit (i.e., 1000 calories) increase in the accumulated caloric deficit, we predict .29 pounds of weight loss.

Using this equation, we can plug in any value of x and obtain the predicted value of y .

$$\hat{y}_i = -.12 + .29x_i$$

ID#124



$$\hat{y}_i = -.12 + .29(14.36) = 4.11$$

id	lbslost	caldef	\hat{y}_i	e_i	e_i^2
101	4.71	9.07	2.55	2.16	4.65
102	2.78	9.30	2.62	0.16	0.03
103	4.93	17.58	5.06	-0.13	0.02
.					
124	7.29	14.36	4.11	3.18	10.11
.					
.					
.					
199	3.18	6.55	1.81	1.37	1.88
200	5.06	19.31	5.57	-0.51	0.26

Besides obtaining the best fit line and corresponding equation, we also want to obtain some additional estimates to gain an understanding of how well this model fits the data. To do this, we will obtain three quantities: Sum of Squares Regression (SSR), Sum of Squares Error/Residual (SSE), and Sum of Squares Total (SST).

$$SST = \sum_{j=1}^N (y_j - \bar{y})^2 = SSR + SSE$$

$$SSR = \sum_{j=1}^N (\hat{y}_j - \bar{y})^2 = SST - SSE$$

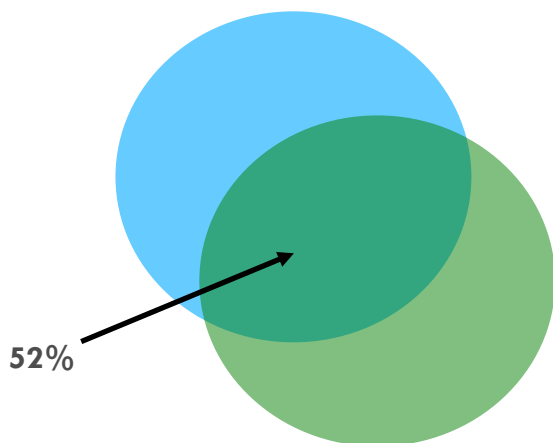
$$SSE = \sum_{j=1}^N (y_j - \hat{y}_j)^2 = \sum_{j=1}^N e_j^2 = SST - SSR$$

id	lbslost	caldef	\hat{y}_i	$(\hat{y}_i - \bar{y})^2$	$(y_i - \hat{y})^2$	$(y_i - \bar{y})^2$
101	4.71	9.07	2.55	0.26	4.65	2.71
102	2.78	9.30	2.62	0.20	0.03	0.08
103	4.93	17.58	5.06	3.99	0.02	3.50
104	6.25	9.19	2.59	0.23	13.40	10.14
105	1.73	8.70	2.44	0.38	0.50	1.77
.						
.						
.						
200	5.06	19.31	5.57	6.28	0.26	3.97
Mean	3.06	10.80				
Σ				227.03	208.80	435.83

We can use these sources of variance to determine how much of the variance in y (pounds lost) can be predicted by caloric deficit. We call this value R^2 (i.e., R-squared).

$$R^2 = \frac{SSR}{SST} = \frac{227.03}{435.83} = .52$$

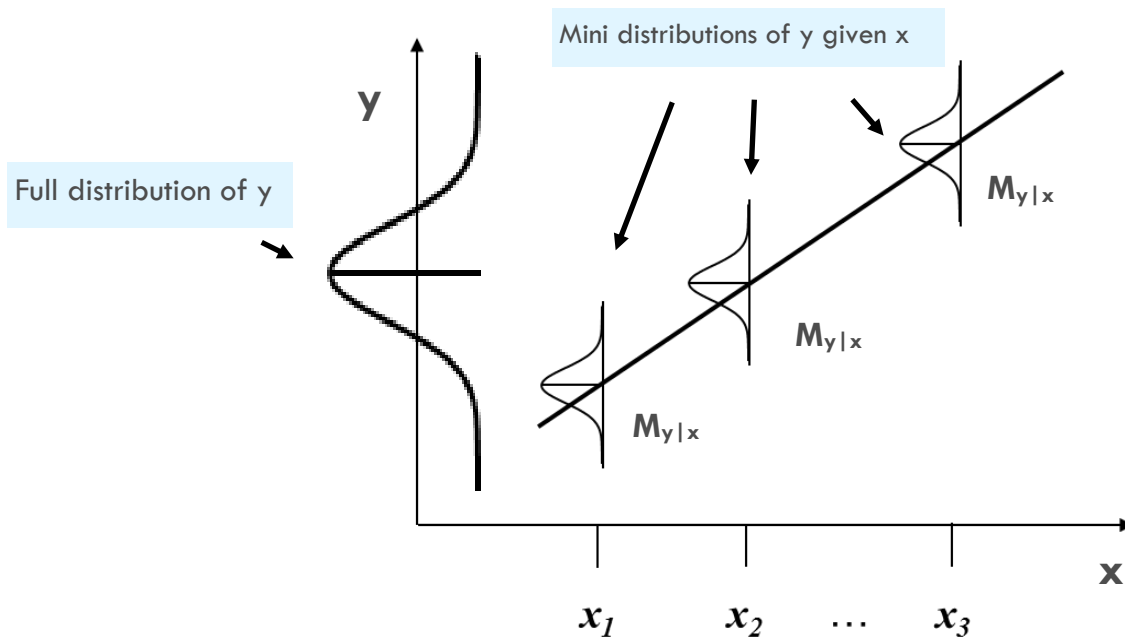
About 52% of the variance in pounds lost can be predicted by caloric deficit.



Our model indicates that 52% of the variance in pounds lost can be explained by caloric deficit. What about the remaining 48%? The following sources of variation are captured by the 48%:

1. Unmeasured or excluded covariates (e.g., genetics, types of food eaten).
2. Measurement error (e.g., scale wasn't perfectly calibrated, measurement of caloric deficit may be imprecise).
3. Random error and individual variation.

Root Mean Square Error



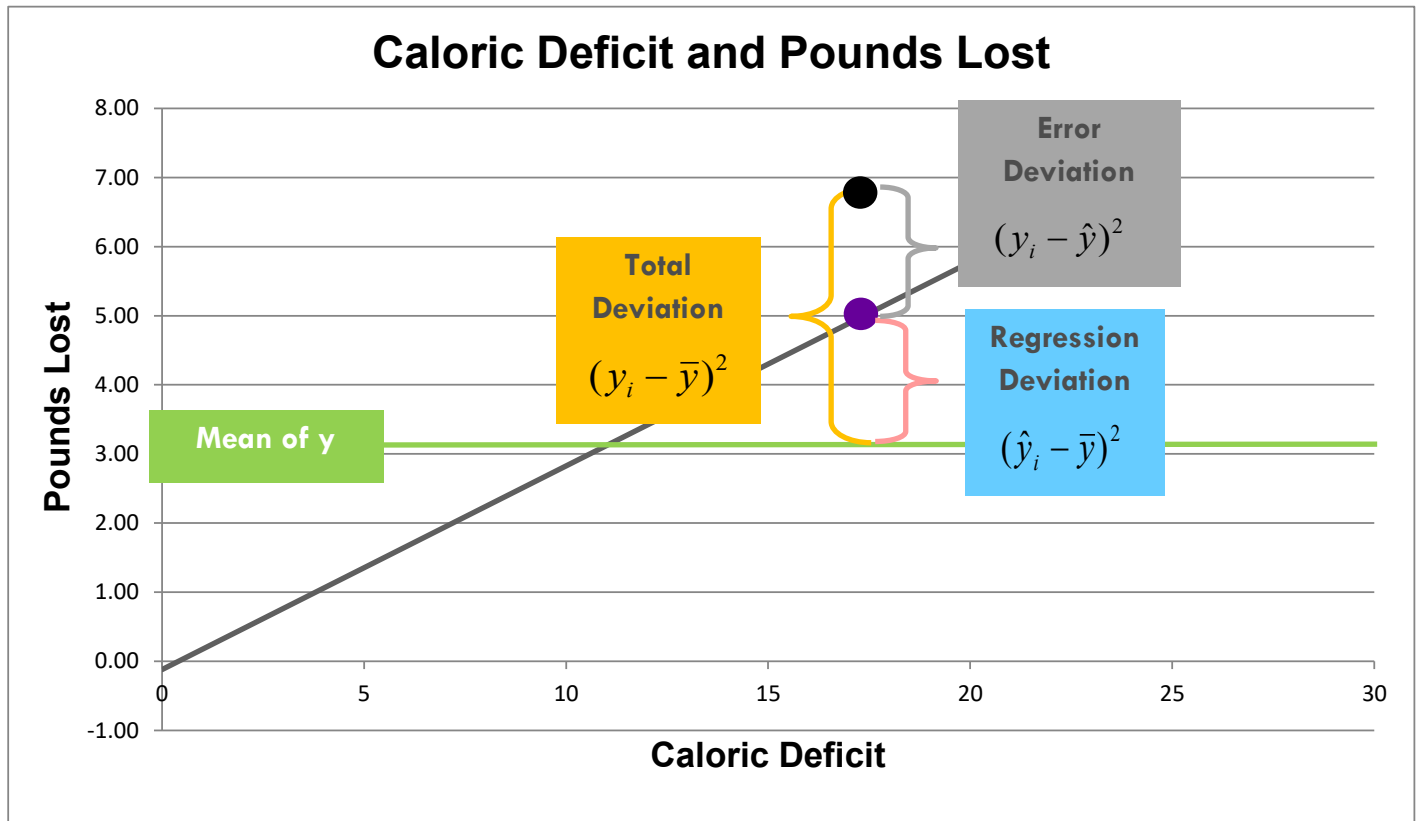
Standard deviation of y

$$\rightarrow SD_y = \sqrt{\frac{\sum (y - \bar{y})^2}{n-1}} = \sqrt{\frac{770.30}{99}} = 2.10$$

Standard deviation of y given
x (written as $y|x$)

$$\rightarrow SD_{y|x} = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}} = \sqrt{\frac{208.80}{98}} = 1.46$$

The Root Mean Square Error (RMSE) is the standard deviation of the residuals and it can be compared to the standard deviation of y. We hope that, after accounting for x, the standard deviation of the residuals is smaller than the standard deviation of y (i.e., because x tells us something about y).



Let's Fit the Model in R

Fit Simple Linear Regression in R

Creates an object that houses the results of the model specified. →

R command for linear regression. Tells R to regress lbslost (y) on caldef (x). Indicates which dataset to use.

```
mod1 <- lm(lbslost ~ caldef, data=wtloss)
```

Requests the results. ←

```
ols_regress(mod1)
```

Model Summary

R	0.722	RMSE	1.460
R-Squared	0.521	Coef. Var	47.647
Adj. R-Squared	0.516	MSE	2.131
Pred R-Squared	0.504	MAE	1.149

RMSE: Root Mean Square Error

MSE: Mean Square Error

MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	227.034	1	227.034	106.558	0.0000
Residual	208.801	98	2.131		
Total	435.835	99			

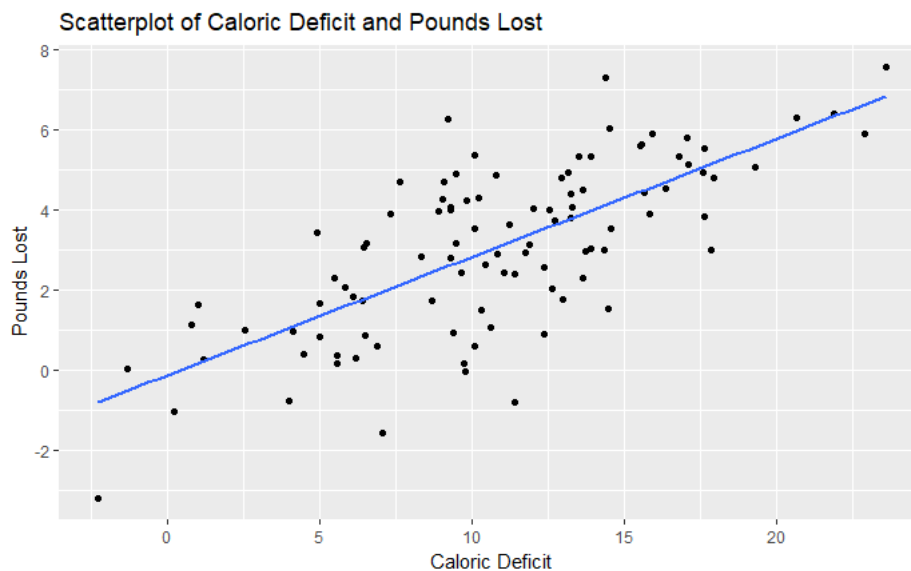
Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-0.121	0.341		-0.354	0.724	-0.798	0.556
caldef	0.295	0.029	0.722	10.323	0.000	0.238	0.351

Intercept and Slope of Best Fit Line

Scatterplot with Best Fit Line

```
ggplot(wtloss, aes(x = caldef, y = lbslost)) +  
  geom_point() +  
  geom_smooth(method = "lm", se = FALSE) +  
  labs(title = "Scatterplot of Caloric Deficit and Pounds Lost", x = "Caloric Deficit", y = "Pounds Lost")
```



Predicted/Fitted Values

We can add the predicted (also called fitted or yhat values) and the residuals (also called errors) to our dataframe.

```
wtloss <- wtloss %>%
  add_predictions(mod1) %>%
  add_residuals(mod1)
```

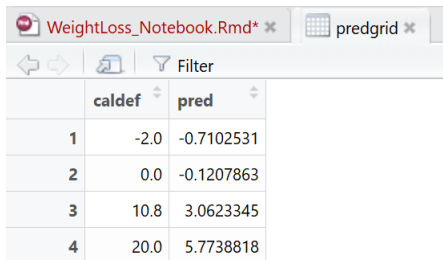
WeightLoss_Notebook.Rmd* x wtloss x					
Filter					
	id	lbslost	caldef	pred	resid
1	101	4.7100912	9.0740314	2.55363391	2.15645729
2	102	2.7846008	9.3019157	2.62079903	0.16380177
3	103	4.9333754	17.5788634	5.06029200	-0.12691660
4	104	6.2482989	9.1908572	2.58806638	3.66023252
5	105	1.7348242	8.7029005	2.44424924	-0.70942504
6	106	4.4411348	15.6525264	4.49253614	-0.05140134
7	107	2.0663362	5.8286340	1.59710689	0.46922931
8	108	0.2984090	6.1675425	1.69699454	-1.39858554
9	109	0.9464819	9.3941337	2.64797875	-1.70149685
10	110	3.7392596	12.6976745	3.62164258	0.11761702
11	111	2.8268606	8.3426084	2.33805912	0.48880148
12	112	0.8721216	6.4925105	1.79277347	-0.92065187
13	113	7.5647785	23.5721318	6.82670839	0.73807011
14	114	1.6656433	5.0076145	1.35512501	0.31051829
15	115	5.6117286	15.5825936	4.47192460	1.13980400
16	116	2.5588428	12.3559142	3.52091440	-0.96207160
17	117	4.8905456	9.4614989	2.66783353	2.22271207
18	118	3.6442292	11.2364945	3.19098402	0.45324518
19	119	0.0417475	-1.3050408	-0.50542537	0.54717287
20	120	3.1529671	9.4578533	2.66675905	0.48620805
21	121	0.2664453	1.2062974	0.23474989	0.03169541
22	122	6.3813312	21.8911868	6.33127775	0.05005345
23	123	2.3140981	13.6566927	3.90429728	-1.59019918
24	124	7.2924021	14.3629003	4.11244025	3.17996185
25	125	5.8065926	17.0513735	4.90482310	0.90176950
26	126	4.2402269	9.8169400	2.77259389	1.46763301
27	127	2.4486352	11.0472418	3.13520493	-0.68656973
28	128	5.9093446	22.8848039	6.62412990	-0.71478530

← This is the person we took a close look at earlier.

Predicted Values for New Cases

We can use `data_grid` to calculate predicted values for new cases as well.

```
predgrid <- data_grid(wtloss, caldef = c(-2, 0, 10.8, 20)) %>%  
  add_predictions(mod1)
```



	caldef	pred
1	-2.0	-0.7102531
2	0.0	-0.1207863
3	10.8	3.0623345
4	20.0	5.7738818

Based on our model, we predict that the average pounds lost among men who accumulate an average caldef (10.8, which is almost 11,000 calories) will lose just a little more than 3 pounds.

Recenter Caloric Deficit to Shift Intercept

By centering x at a different value, we can define the intercept of the equation to represent the predicted value of y for any group we want. For example, we could center at the mean of x (i.e., $x - 10.80$) or at 3500 (i.e., $x - 3.5$). This is particularly useful when a score of 0 on x isn't possible or is not in the range of observed scores.

```
wtloss <- wtloss %>%
  mutate(caldef_m = caldef - mean(caldef))

mod2 <- lm(lbslost ~ caldef_m, data = wtloss)
ols_regress(mod2)
```

The first part of this syntax creates a new variable in the wtloss data set. We name it caldef_m, and it is defined as the original caldef score minus the mean caldef in the sample (10.8).

Model Summary

R	0.722	RMSE	1.460
R-Squared	0.521	Coef. Var	47.647
Adj. R-Squared	0.516	MSE	2.131
Pred R-Squared	0.504	MAE	1.149

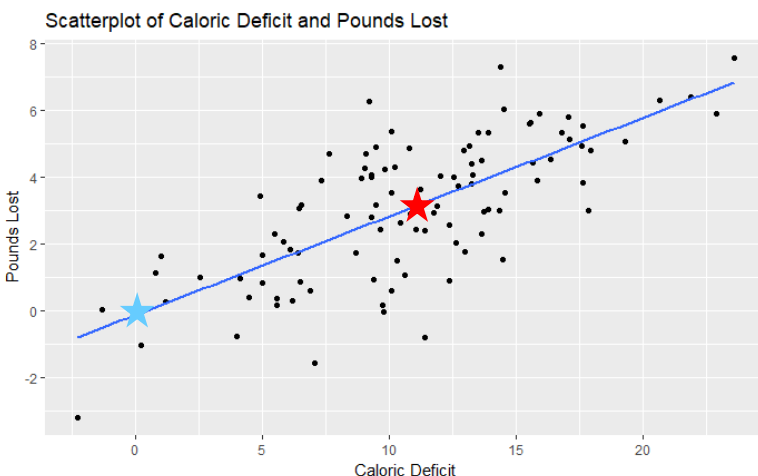
RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	227.034	1	227.034	106.558	0.0000
Residual	208.801	98	2.131		
Total	435.835	99			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig.	lower	upper
(Intercept)	3.063	0.146		20.988	0.000	2.774	3.353
caldef_m	0.295	0.029	0.722	10.323	0.000	0.238	0.351



Re-centering a predictor (i.e., subtracting a constant) doesn't change the overall model fit, it just shifts the location of the intercept.

Inference for SLR

Population Parameters

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

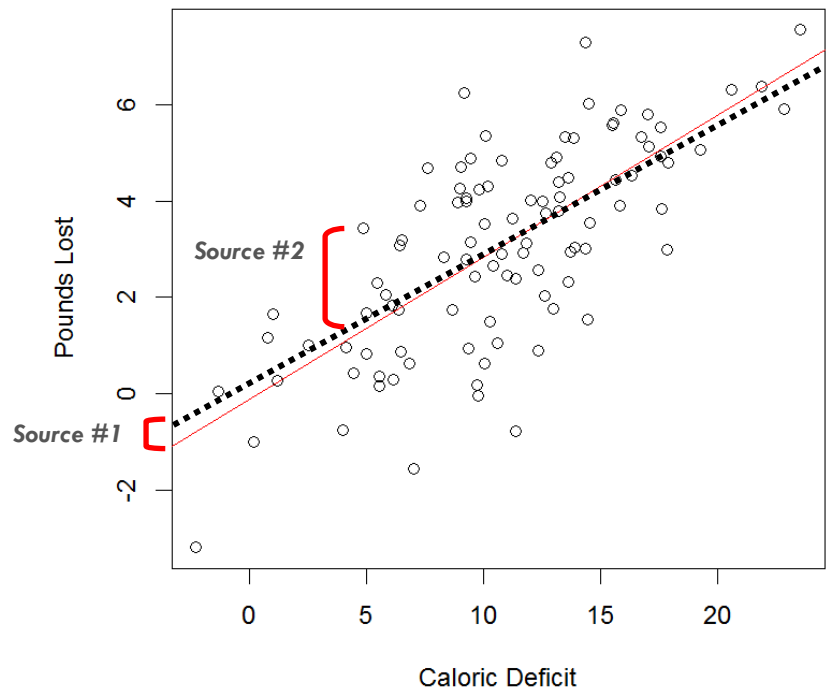
Estimated Parameters

$$Y_i = b_0 + b_1 X_i + e_i$$

The population parameters (β_0 , β_1 , and ε_i) are unknown. We use a randomly selected sample to estimate these parameters.

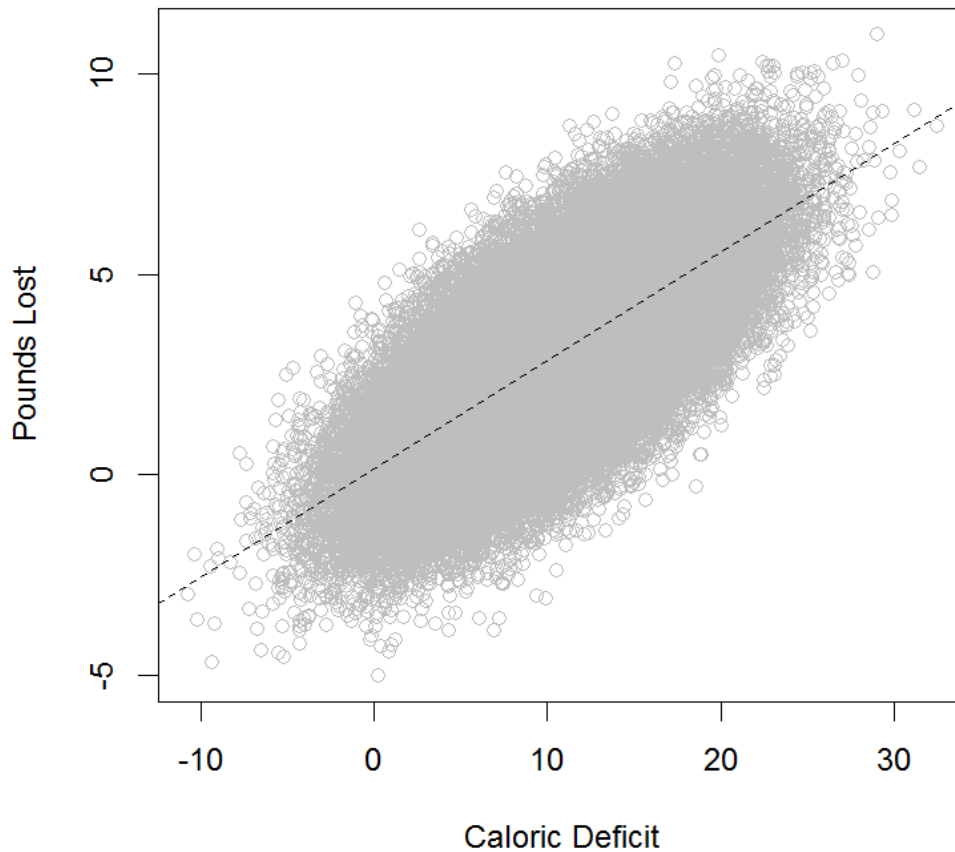
Imagine that the black dotted line represents the true relationship between x and y . In our single sample (represented by the red line), we find a line that is close but not exact.

Scatterplot of Caloric Deficit and Pounds Lost



Sources of error in the prediction of an outcome (y)

1. Error because the true and predicted lines are not the same (i.e., the black dotted line that represents the “true effect in the population” and our red estimated regression line from the sample do not overlap—see SOURCE 1 above).
2. Error because the model is not deterministic – we can’t perfectly predict y from x (i.e., we have a residual to account for individual variation—see SOURCE 2 above).

Caloric Deficit and Pounds Lost - Population

The population for our caloric deficit and weight loss example is all obese men in Northern Colorado. Let's imagine the researcher recruited every single obese man in Northern Colorado to participate in the study. In this case, we would have a census, i.e., the entire population, and we therefore would know the true values of β_0 and β_1 .

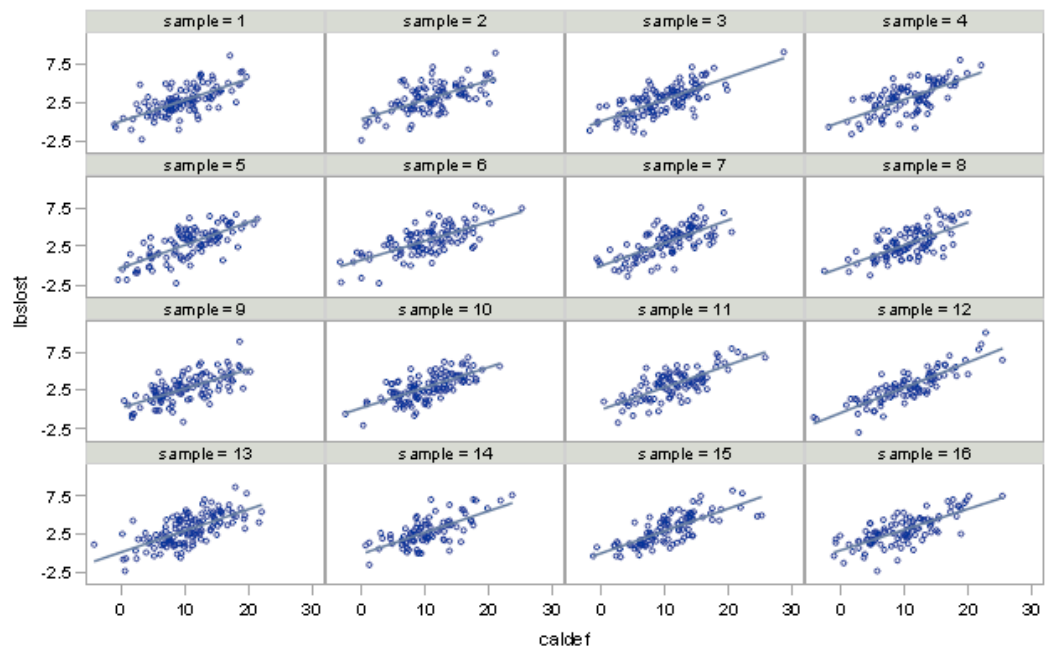
$$\text{lbslost}_i = .15 + .27(\text{caldef}_i) + \varepsilon_i \quad \leftarrow$$

POPULATION EQUATION

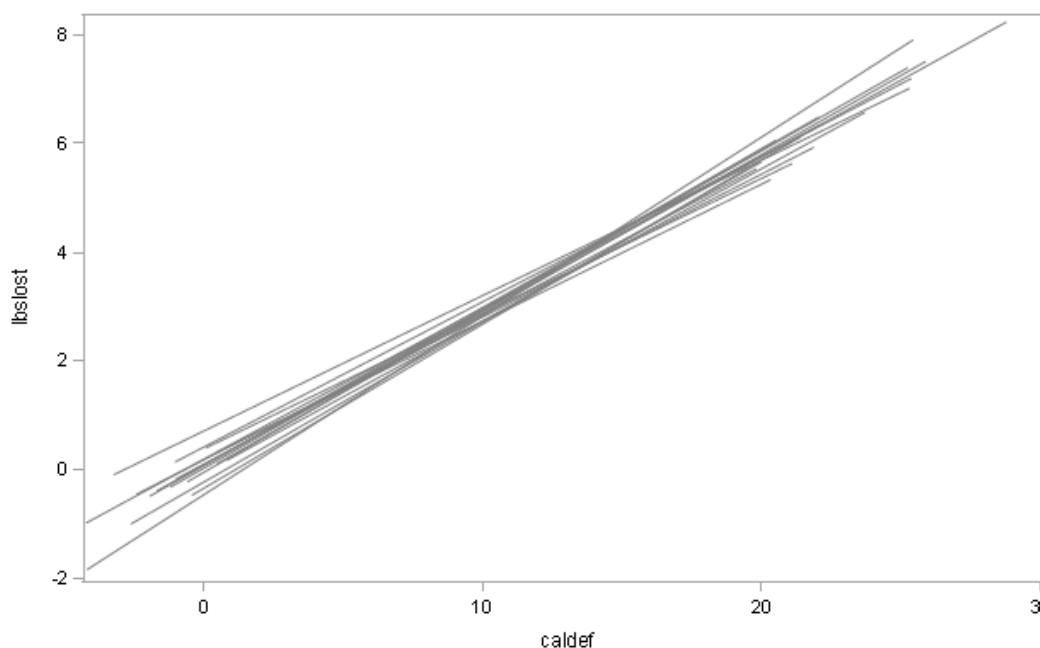
Imagine that we Drew Many, Many Random Samples from the Population—Here are the Results for 16 Samples

Scatter plot of caloric deficit & pounds lost

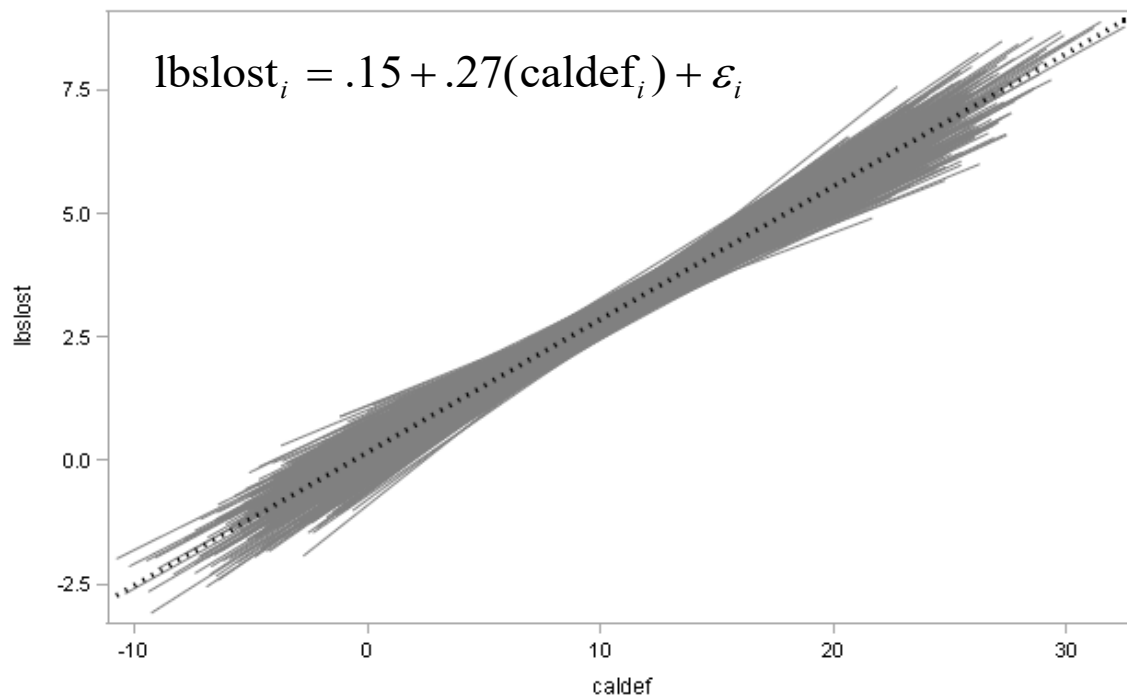
Sample	b_0	b_1
1	0.10	0.27
2	0.37	0.25
3	0.08	0.28
4	0.06	0.29
5	-0.34	0.30
6	0.71	0.25
7	-0.05	0.30
8	-0.24	0.29
9	0.20	0.25
10	0.17	0.26
11	-0.05	0.29
12	-0.47	0.33
13	0.18	0.28
14	-0.07	0.28
15	0.02	0.29
16	0.41	0.27



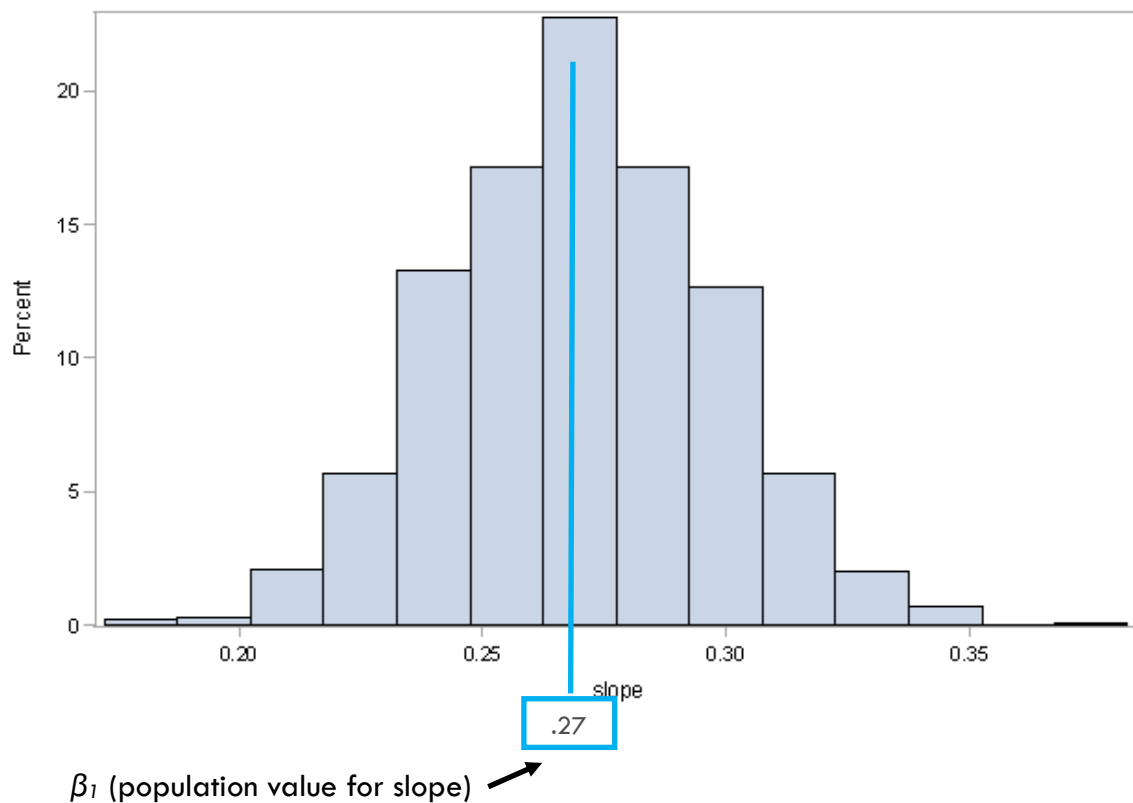
Let's plot all 16 best fit lines on one graph:



Regression equation for population (dashed line) & 1000 random samples



Distribution of slope (b_1) across 1000 random samples



Null Hypothesis Tests for Regression Parameters

In the same way that we determined whether the average pounds lost was significantly different from zero using our one-sample t-test, we can use a t-ratio to test if the intercept and slope are each different from zero.

For the intercept, the null and alternative hypotheses are:

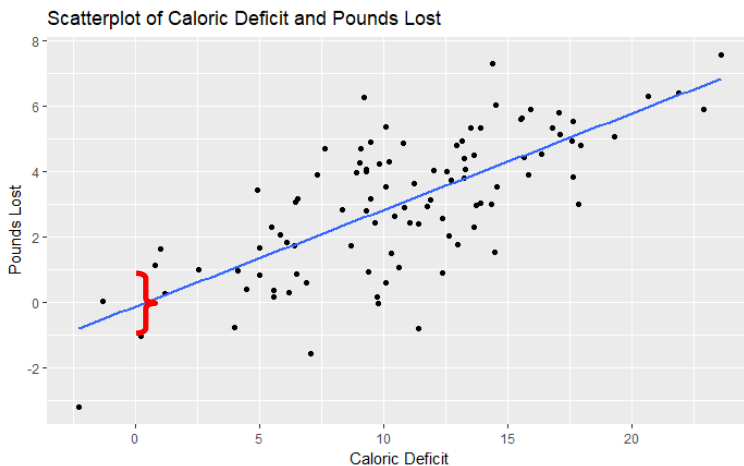
$$H_0 = \beta_0 = 0$$

$$H_a = \beta_0 \neq 0$$

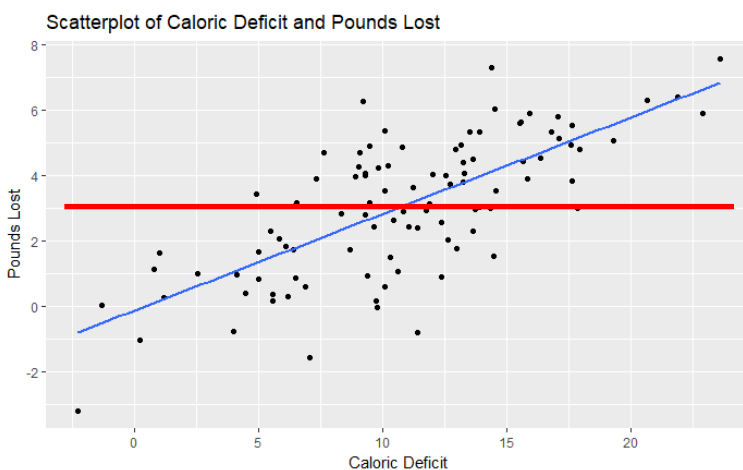
For the slope, the null and alternative hypotheses are:

$$H_0 = \beta_1 = 0$$

$$H_a = \beta_1 \neq 0$$



With the null hypothesis for the intercept, we wonder if 0 is a plausible value for the intercept in the population. This is a very similar question to our one-sample t-test, but here, rather than a simple mean, we're using a conditional mean (the mean when caloric deficit equals 0).



With the null hypothesis for the slope, we wonder if a slope of 0 (a flat line through the mean of y (pounds lost) is plausible. A slope of 0 would mean that caloric deficit tells us nothing about pounds lost, in other words, that caloric deficit is unrelated to pounds lost.

Calculation of Standard Errors for b_0 and b_1

$$se(b_1) = \frac{RMSE}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{1.46}{\sqrt{2613.56}} = .03$$

$$se(b_0) = RMSE \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}} = 1.46 \sqrt{\frac{1}{100} + \frac{10.80^2}{2613.56}} = .34$$

Calculation of t-ratios for b_0 and b_1 for SLR

$$t^* = \frac{b}{se} \sim t_{(n-2)df}$$

~ means "is distributed as"

t^* is compared to the critical t (df equal to $n-1$ -# of predictors). For a two-tailed test and an alpha of .05, critical t for 98 df (i.e., $n-1$ -# of predictors or... $100-2=98$) is 1.985. If the absolute value of t^* exceeds the absolute value of critical t , then we reject the null hypothesis.

`qt(c(.025, .975), df=98)`

[1] -1.984467 1.984467

t^* for intercept (b_0):

$$t^* = \frac{-.12}{.34} = -.35$$

t^* for slope (b_1):

$$t^* = \frac{.29}{.03} = 10.32$$

$|t^*|$ exceeds |critical t | for the slope, but not the intercept. For the intercept, this means that the pounds lost among people who have a 0 caloric deficit is not significantly different from 0 (i.e., 0 is a plausible value). For the slope, this means that the increase in pounds lost (i.e., change in y) for a one unit increase in the caloric deficit is significantly different from 0 (i.e., 0 is not a plausible value).

Let's Return to Our Regression Model Output

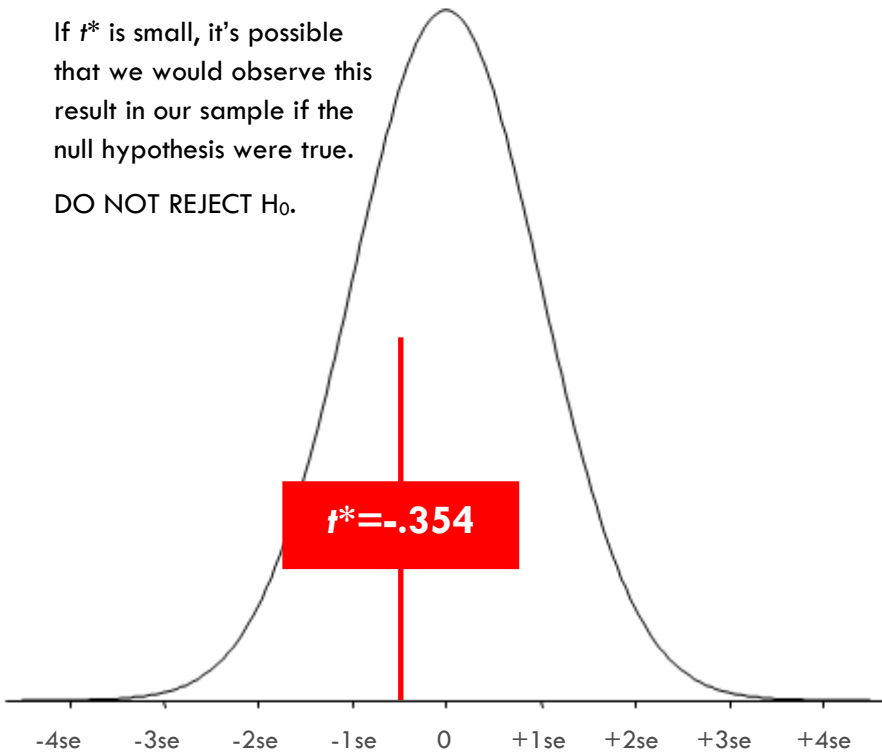
Parameter Estimates							
model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-0.121	0.341		-0.354	0.724	-0.798	0.556
caldef	0.295	0.029	0.722	10.323	0.000	0.238	0.351

The orange box shows the standard error (se). The blue box shows the t^* (labeled simply t) and the p-value (labeled Sig). t^* shows how many standard errors away the regression coefficient (labeled beta) is from 0 and the p-value shows the probability that we would obtain a t^* at this magnitude or larger if the null hypothesis were true.

Sampling distribution for intercept (top) and slope (bottom) under the null hypothesis.

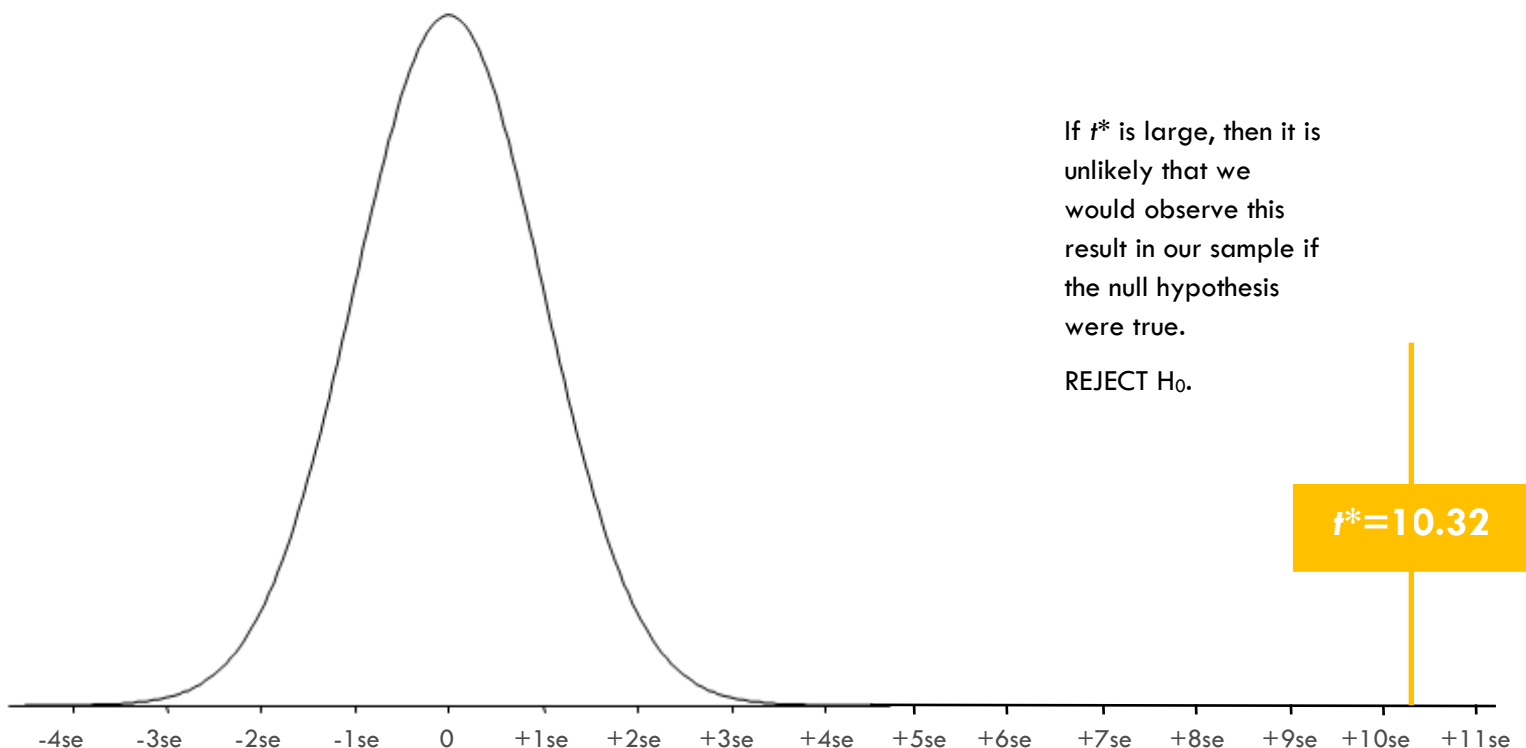
If t^* is small, it's possible that we would observe this result in our sample if the null hypothesis were true.

DO NOT REJECT H_0 .



If t^* is large, then it is unlikely that we would observe this result in our sample if the null hypothesis were true.

REJECT H_0 .



Confidence Intervals (CI)

In the same way that we calculated a CI for the mean in Unit 3, we can calculate a CI for each regression parameter. These display in the regression model output. By default, the CI is a 95% CI. We can use the `confint` function in R to obtain other CIs.

Parameter Estimates							
model	Beta	Std. Error	Std. Beta	t	Sig.	lower	upper
(Intercept)	-0.121	0.341		-0.354	0.724	-0.798	0.556
caldef	0.295	0.029	0.722	10.323	0.000	0.238	0.351

95% Confidence Interval for b_0

$$b_0 \pm t_{crit}[se(b_0)]$$

$$-.12 \pm 1.98(.34)$$

$$(-.80, .56)$$

95% Confidence Interval for b_1

$$b_1 \pm t_{crit}[se(b_1)]$$

$$.29 \pm 1.98(.03)$$

$$(0.24, .35)$$

95% of the 95% confidence intervals we construct will include the population value

```
confint(mod1, level=.95)
confint(mod2, level=.99)
```

```

      2.5 %    97.5 %
(Intercept) -0.7980177 0.5564452
caldef       0.2380728 0.3513940    95% CI
      0.5 %    99.5 %
(Intercept) -1.0172688 0.7756963
caldef       0.2197291 0.3697377    99% CI
```

Notice that the 95% CI for the slope doesn't contain 0. This indicates that 0 is not a plausible value for the slope estimate. But the 95% CI for the intercept does contain 0. This indicates that 0 is a plausible value for the intercept. This corresponds with our hypothesis test for these estimates. The t^* for $\alpha = .05$ will always correspond with the 95% CI for the estimate. If the absolute value of t^* exceeds the absolute value of the t_{crit} , then the 95% CI will not include 0. Otherwise the 95% CI will include 0.

95% CI for the a New y

We can also construct confidence intervals for new predictions. There are two types of possible intervals.

The first constructs a 95% confidence interval for the mean of y given x . First, imagine that we wanted to construct a confidence interval that would give us a range of plausible values for the mean of y (e.g., pounds lost) at a certain value of x (e.g., 7,000 calorie deficit). “Among men who accumulate a caloric deficit of 7,000 calories, what is their predicted average weight loss and how precise is this estimate?”

The second constructs a 95% prediction interval for a single new y . Imagine that we wanted to construct a prediction interval that would give us a range of plausible values for predicting a new y (e.g., pounds lost) at a certain value of x (e.g., 7,000 calorie deficit). “If Peter Griffin accumulates a caloric deficit of 7,000 calories, how much weight do we predict that he will lose?”

For both intervals, we start by solving for \hat{y} given x .

$$\hat{y}_x = -.12 + .29(7.0) = 1.94$$

Let's add this value of interest, 7, to our `data_grid` code that we constructed earlier.

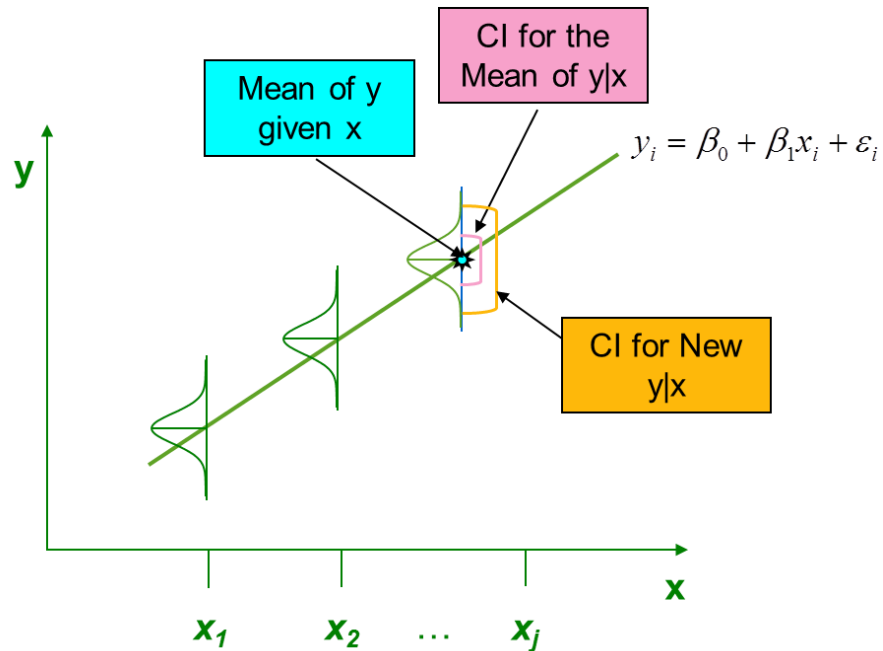
```
# pick a few prototypical values
predgrid <- data_grid(wtloss, caldef = c(-2, 0, 7, 10.8, 20))
predict(mod1, predgrid, interval = "confidence", level = .95)
predict(mod1, predgrid, interval = "prediction", level = .95)
```

	fit	lwr	upr
1	-0.7102531	-1.4914220	0.07091589
2	-0.1207863	-0.7980177	0.55644521
3	1.9423476	1.5812927	2.30340242
4	3.0623345	2.7726686	3.35200042
5	5.7738818	5.1777232	6.37004042

← Confidence intervals for the mean of y at a given x

	fit	lwr	upr
1	-0.7102531	-3.7103954	2.289889
2	-0.1207863	-3.0955588	2.853986
3	1.9423476	-0.9767258	4.861421
4	3.0623345	0.1512290	5.973440
5	5.7738818	2.8165124	8.731251

← Prediction intervals for a new y

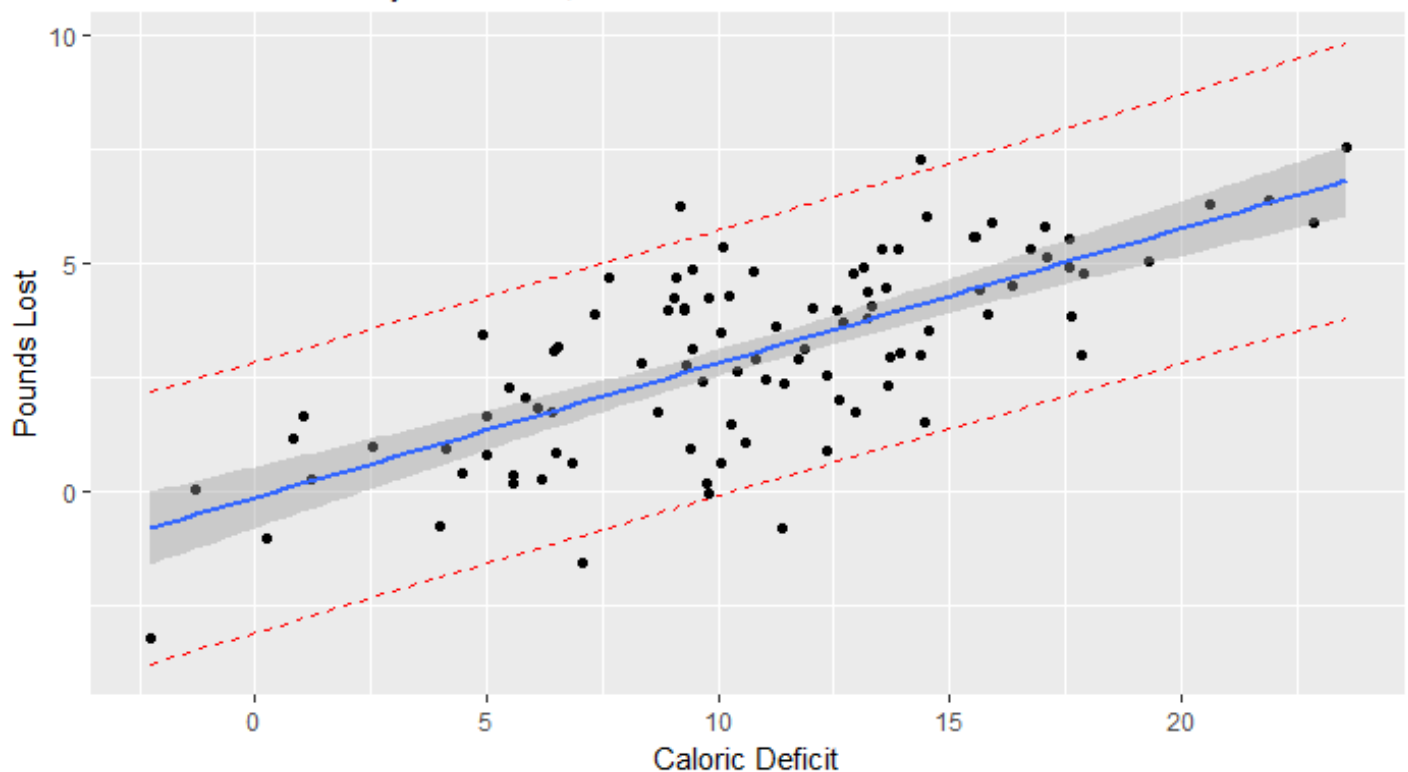


Since both sampling variation and individual variation exist, the width of the prediction interval for a new y will be much wider than that for the mean of y .


```
temp_var <- predict(mod1, wtloss, interval="prediction")  
wtloss_new <- cbind(temp_var, wtloss)  
  
ggplot(wtloss_new, aes(x = caldef, y = lbslost)) +  
  geom_point() +  
  geom_smooth(method = "lm", se = TRUE) +  
  geom_line(aes(y=lwr), color = "red", linetype = "dashed") +  
  geom_line(aes(y=upr), color = "red", linetype = "dashed") +  
  labs(title = "Scatterplot of Caloric Deficit and Pounds Lost with Best Fit Line",  
        subtitle = "Confidence Interval in Gray Shaded Area, Prediction Interval in Red",  
        x = "Caloric Deficit", y = "Pounds Lost")
```

Scatterplot of Caloric Deficit and Pounds Lost with Best Fit Line

Confidence Interval in Gray Shaded Area, Prediction Interval in Red



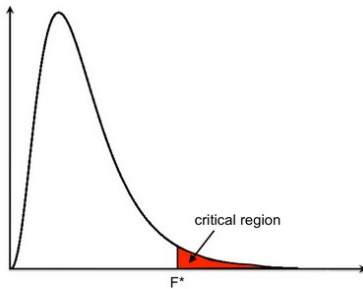
F-test and SLR

In a SLR, we can equivalently test the relationship between two variables (i.e., x predicting y) through a F-test.

$H_0: \beta_1 = \beta_2 = \beta_k = 0$ (the regression coefficient for all predictors is 0, no variance in the outcome is explained by the predictors)

H_a : not all $\beta_k = 0$ (the predictors predict variability in the outcome)

Because we have only 1 predictor in a SLR, this tests the same quantity at the null hypothesis for the slope.



A test for the ratio of two variances requires the F-distribution. Like the Student's t , there are a family of F distributions corresponding to the numerator and denominator df . The df for the numerator equals the # of predictors; the df for the denominator equals $n - 1 - \#$ of predictors. F-distributions are generally skewed.

Critical value of F (F_{crit}) `qf(.95, df1=1, df2=98)` [1] 3.938111

$$F^* = \frac{SSR/df}{SSE/df} = \frac{MSR}{MSE} = \frac{227.03/1}{208.80/98} = \frac{227.03}{2.13} = 106.56$$

Critical F ($\alpha = .05$) for this problem is 3.94. F^* exceeds F_{crit} , therefore, reject H_0 .

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	227.034	1	227.034	106.558	0.0000
Residual	208.801	98	2.131		
Total	435.835	99			

This is the F^*

↖

The probability that we would achieve an F-value of this magnitude or larger if the null hypothesis were true.

`pf(106.558, df1=1, df2=98, lower.tail = FALSE)`

Note that t^* squared from the estimate of the slope equals F^* . $10.32^2 = 106.56$. F-tests become much more useful/interesting in MLR...stay tuned.

Pearson's Correlation Coefficient (r)

$$r_{xy} = \frac{\sum z_x z_y}{n-1} = \frac{71.45}{99} = .72$$

where z_x & z_y are standardized scores (mean of 0 and standard deviation of 1) of x and y respectively – “z-scores.”

id	lbslost	caldef	z_y	z_x	$z_y z_x$
101	4.71	9.07	0.78	-0.34	-0.26
102	2.78	9.30	-0.13	-0.29	0.04
103	4.93	17.58	0.89	1.32	1.18
104	6.25	9.19	1.52	-0.31	-0.48
105	1.73	8.70	-0.63	-0.41	0.26
.					
.					
.					
200	5.06	19.31	0.95	1.65	1.57
Mean	3.06	10.80			
SD	2.10	5.14			
Σ					71.45

Some properties of r :

- r is independent of the units of measurement
- r varies between 1 (perfect positive r) and -1 (perfect negative r), when $r = 0$ there is no relationship
- The sign of r indicates the direction of the relationship. A positive sign indicates that higher values of x are associated with higher values of y and a negative sign indicates that higher values of x are associated with smaller values of y
- r is called a zero-order correlation
- In a SLR, the square root of R^2 equals the absolute value of the r (e.g., square root of .52 equals .72)
- In a SLR, the absolute value of the r equals the correlation between the observed y and predicted y
- The slope of a SLR in which the zscore of y is regressed on the zscore of x will equal the correlation between y and x , and this is the definition of the standardized beta in the parameter estimates table below.
- You can calculate a correlation coefficient (r) from the unstandardized regression slope (b_1) and the standard deviation (sd) of x and y :

$$r_{xy} = \frac{sd_x}{sd_y} b_1 = \frac{5.14}{2.10} \cdot .29 = .72$$

`cor.test(wtloss$caldef, wtloss$lbslost)`

Pearson's product-moment correlation

```
data: wtloss$caldef and wtloss$lbslost
t = 10.3227, df = 98, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6121017 0.8041627
sample estimates:
      cor
0.7217459
```

$$t^* = \frac{r_{xy}}{\sqrt{(1-r_{xy}^2)/(n-2)}} = \frac{.72}{\sqrt{(1-.72^2)/98}} = 10.32$$

produces same value as t^ for the unstandardized slope (b_1).

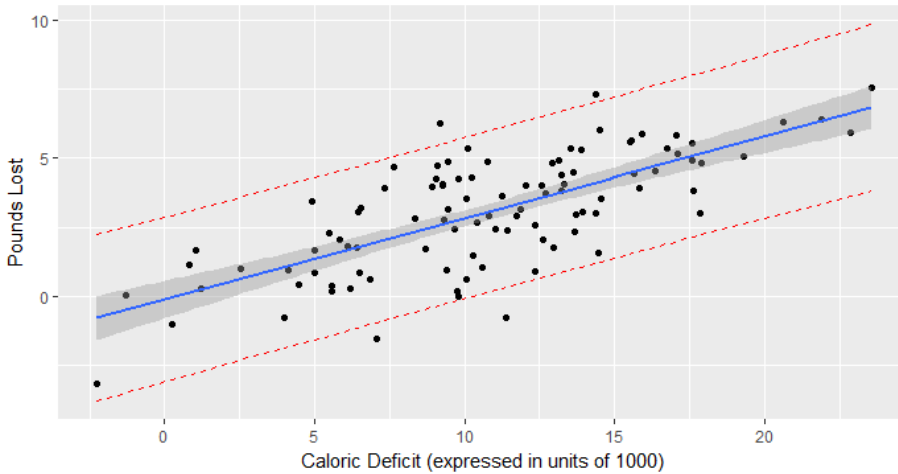
Parameter Estimates							
model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-0.121	0.341		-0.354	0.724	-0.798	0.556
caldef	0.295	0.029	0.722	10.323	0.000	0.238	0.351

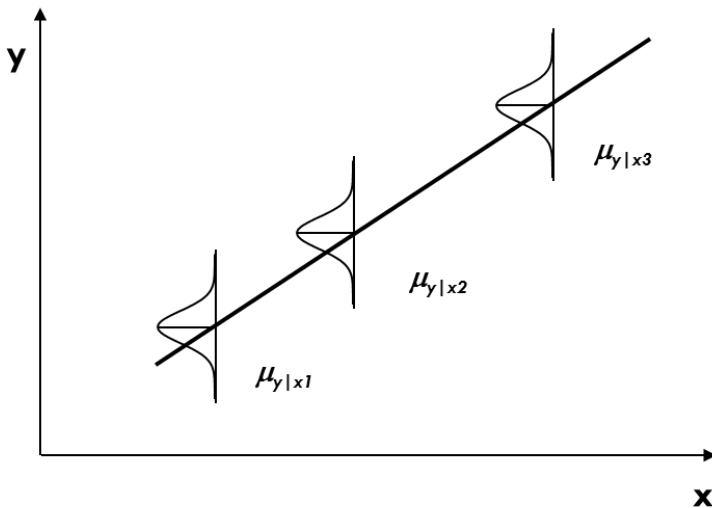
Write Up

We sought to estimate the relationship between caloric deficit and pounds lost. One hundred obese men from Northern Colorado were randomly selected. They were each instructed to follow a weight loss program over the course of one month. Total pounds lost on the program was calculated by subtracting ending weight from starting weight. Caloric deficit was calculated each day by subtracting caloric intake from caloric expenditure. The caloric deficit was summed across all days to form the total for the month and then divided by 1000. Using a simple linear regression, pounds lost was regressed on caloric deficit. Approximately 52% of the variance in pounds lost was predicted by caloric deficit. Each one unit increase in caloric deficit (1000 calories) was associated with .29 pounds lost ($b = .29$, $t(98) = 10.32$, 95% CI .24, .35). The results of the model are depicted in Figure 1.

Figure 1
Scatterplot of Caloric Deficit and Pounds Lost with Best Fit Line

95% Confidence Interval in Gray Shaded Area, 95% Prediction Interval in Red



Assumptions of Fitting an OLS Regression Line

1. At each value of x , there is a distribution of y . These distributions have a mean $\mu_{y|x}$ and a variance of $\sigma^2_{y|x}$.

2. The relationship between x and y is linear. The means of each of these distributions, the $\mu_{y|x}$'s, may be joined by a straight line.

3. Homoscedasticity. The variances of each of these distributions, the $\sigma^2_{y|x}$'s, are equivalent.

4. Independence of observations. At each given value of x (at each x_i), the values of y (y_i 's) are independent of each other.

5. Normality – for inference only
At each given value of x (at each x_i), the values of y (the y_i 's) are normally distributed.

Note, $\sigma^2_{y|x}$ means “the variance of y given (or controlling for) x .”

Assessment of these assumptions is pretty simple in SLR—a scatterplot is really all you need. But, it becomes more difficult to assess assumptions when the regression model becomes more complex (e.g., multiple predictors, polynomial terms, interactions). Unit 7 is entirely dedicated to learning about how we can use R to help us evaluate the assumptions, and find a remedy if an assumption is violated.