Chapter 1

STAT 3240

Michael McIsaac

UPEI

1: Linear Regression with One Predictor Variable

Regression analysis is widely used in business, the social and behavioural sciences, the biological sciences, and many other disciplines.

The goal is to predict a *response* or outcome variable from other *predictor* variables. E.g.,

1. Sales of a product can be predicted by utilizing the relationship between sales and amount of advertising expenditures.
2. The performance of an employee on a job can be predicted from a battery of aptitude tests.
3. The size of the vocabulary of a child can be predicted from the age of the child and amount of education of the parents.
4. The length of hospital stay of a surgical patient can be predicted from the severity of the operation.

In Part I, we focus on regression analysis with a single predictor variable is used.

In Part II, we focus on regression with multiple predictors.

A few of my projects:

- Predicting organized sport participation among youth from family structure
- Predicting substance-use behavior from involvement in team sport culture in Canadian adolescents
- Family structure as a predictor of screen time among youth
- Distance to specialist medical care and diagnosis of obstructive sleep apnea in rural Saskatchewan

Learning Objectives for Sections 1.1, 1.2, 1.4

After Sections 1.1, 1.2, 1.4, you should be able to

- Describe the uses of regression analysis
- Contrast regression vs causation
- Identify observational and experimental data and contrast these with respect to causation

1.1: Relations between Variables

A functional relationship between two variables is expressed by a mathematical formula:

$$Y = f(X),$$

where

- Y represents the *dependent variable* and
- X represents the *independent variable*.

For example, consider the relation between dollar sales (Y) of a product sold at a fixed price and number of units sold (X). If the selling price is $2 per unit, the relation is expressed by the equation

$$Y = 2 \cdot X.$$

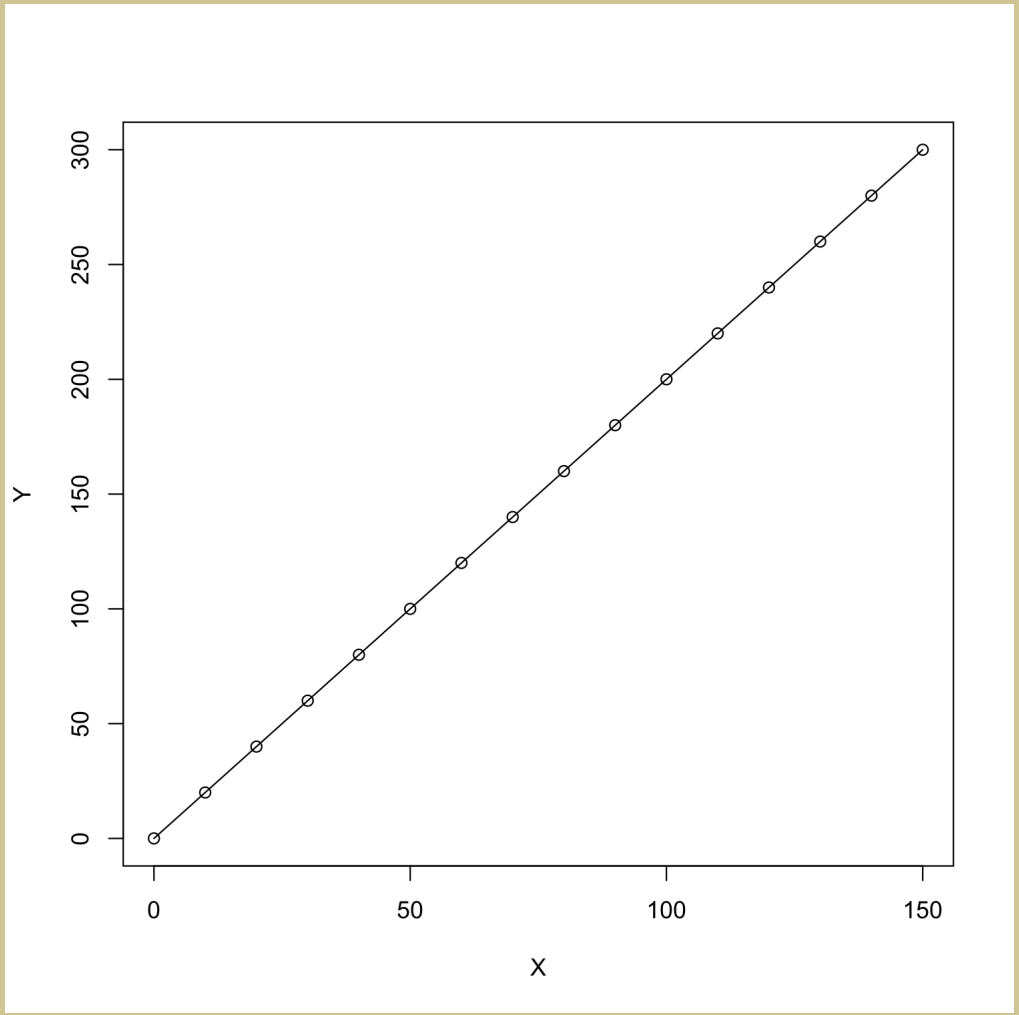
```r
X = seq(0, 150, by=10)
Y = 2*X
data.frame(units=X, sales=Y) %>% datatable()
```

Show 20 entries Search:

	units	sales
1	0	0
2	10	20
3	20	40
4	30	60
5	40	80
6	50	100
7	60	120

Showing 1 to 16 of 16 entries Previous 1 Next

```
plot(X,Y, type="l")
points(X,Y)
```



Notice that all observations fall directly on the line

A functional relationship does not need to be *linear*:

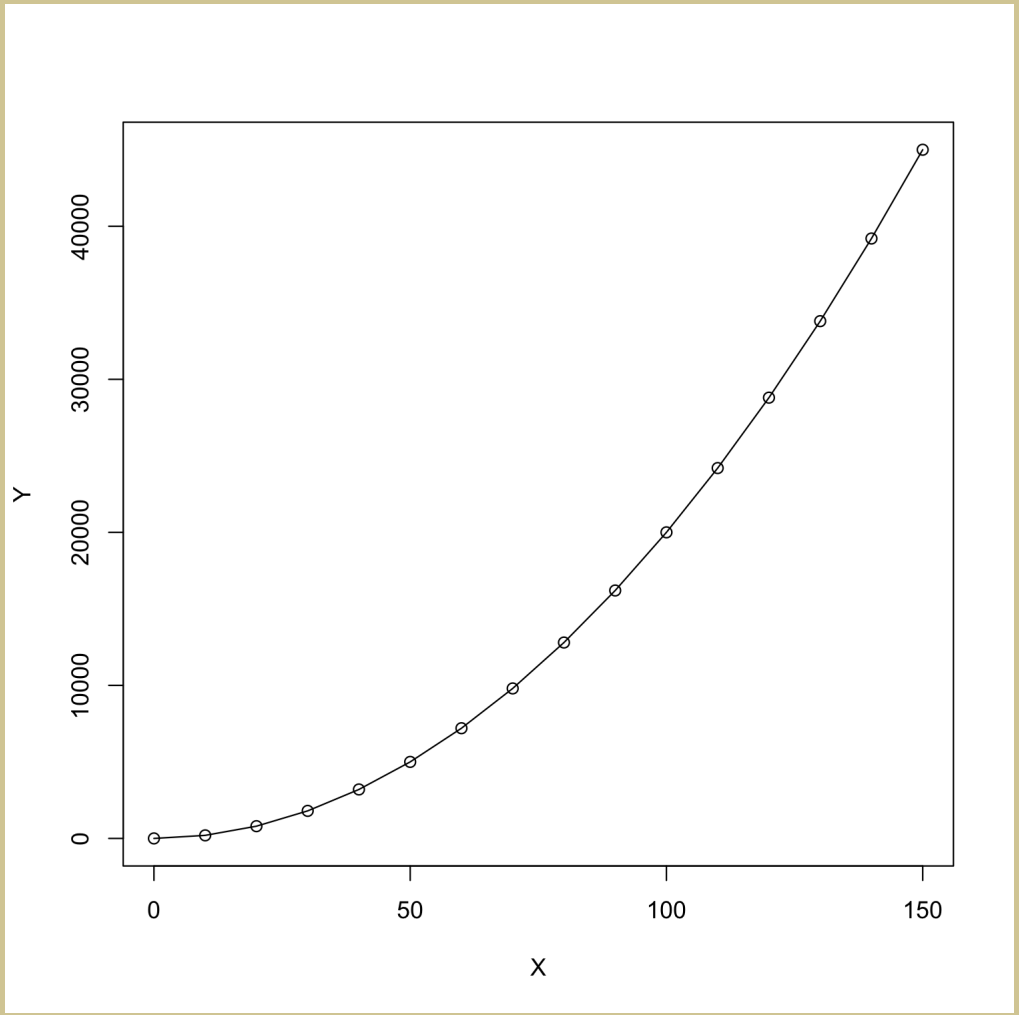
```
X = seq(0, 150, by=10)
Y = 2*X^2
data.frame(units=X, sales=Y) %>% datatable()
```

Show 20 entries Search:

	units	sales
1	0	0
2	10	200
3	20	800
4	30	1800
5	40	3200
6	50	5000
7	60	7200

Showing 1 to 16 of 16 entries Previous 1 Next

```
plot(X,Y, type="l")
points(X,Y)
```



This is an example of a *curvilinear* relationship.

Statistical Relation between Two Variables

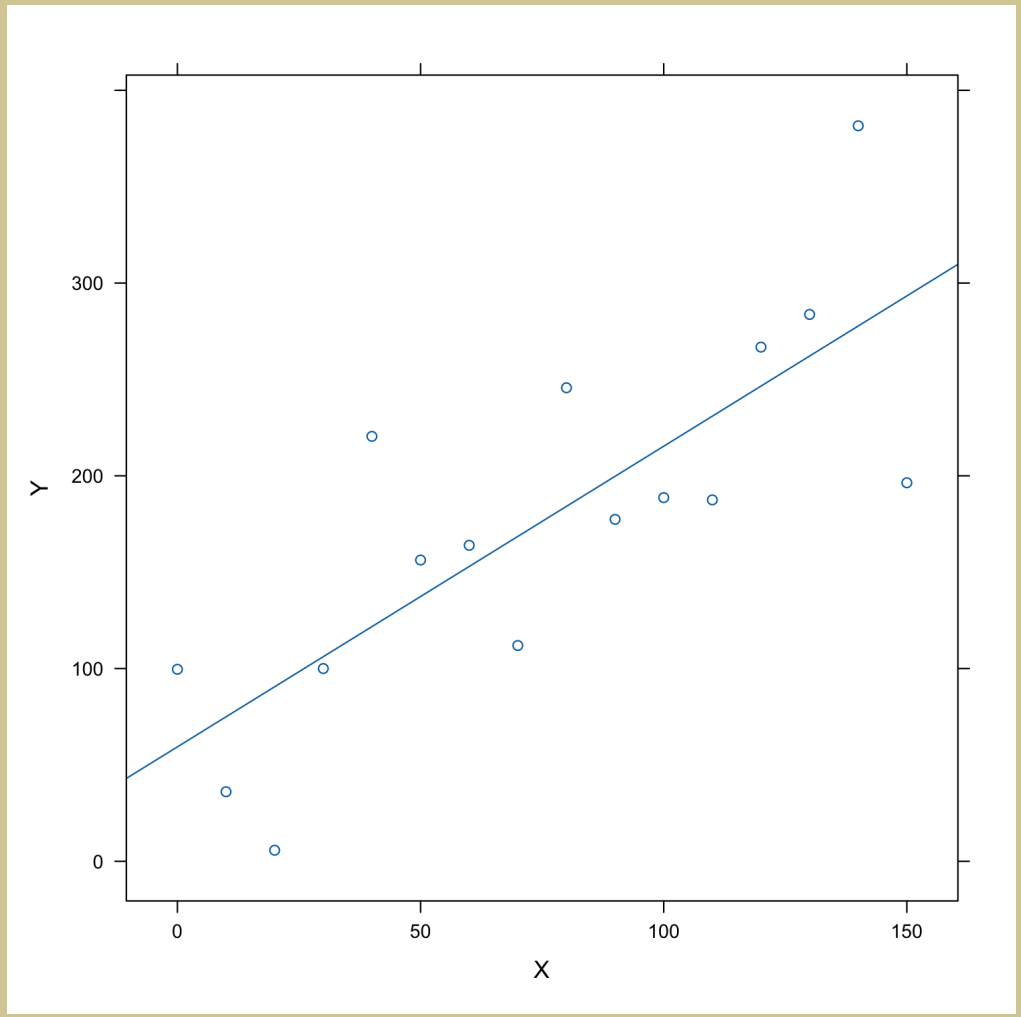
A statistical relation, unlike a *functional relation*, is not a perfect one.

- In general, the observations from a statistical relation do not fall directly on the underlying curve.

For example, consider now the relationship between dollar sales (Y) of a product sold at a fixed price and the amount of money spent on advertising (X).

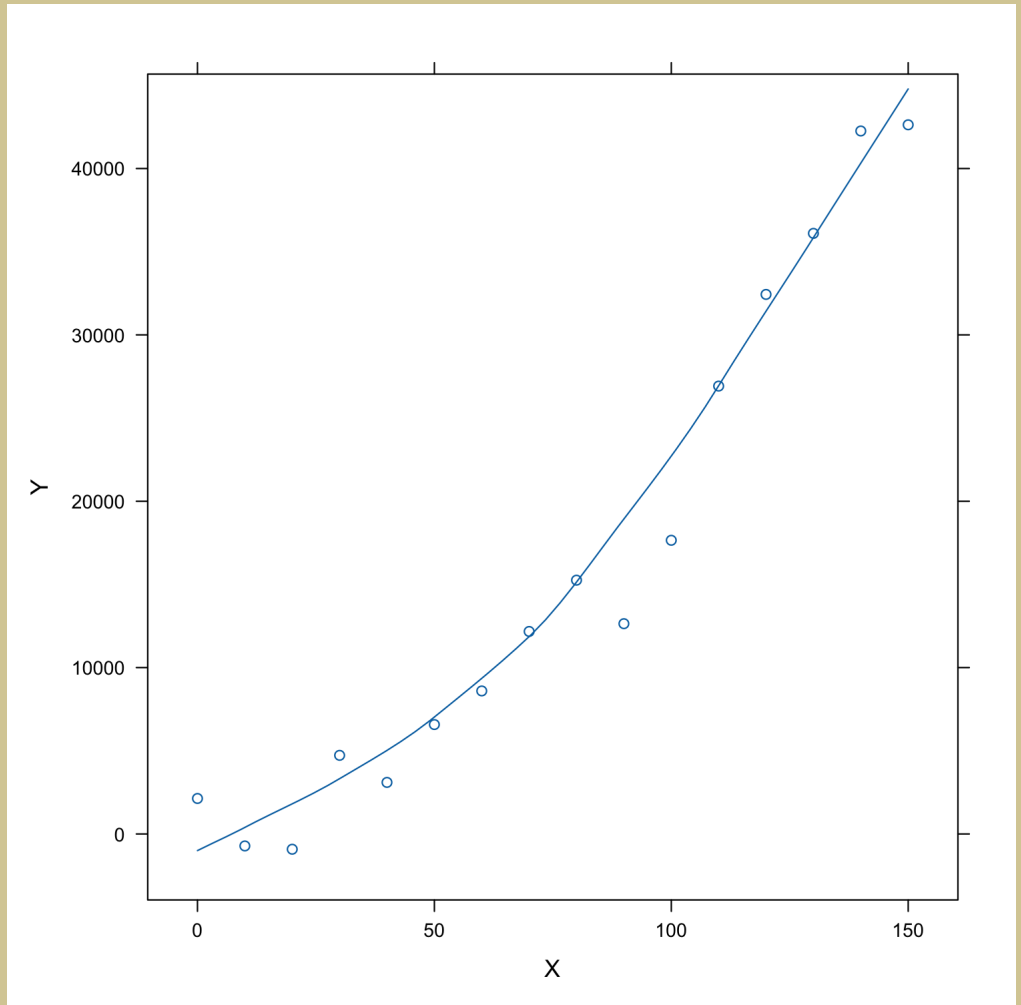
As X increases, you would generally expect Y to increase, but you certainly wouldn't expect every dollar in advertising to give you the exact same increase in sales.

```
X = seq(0, 150, by=10)
Y = 2*X + rnorm(length(X), 0, 50)
xyplot(Y~X,  type=c("p", "r"))
```



It is, of course, possible to have *curvilinear* statistical relations:

```
X = seq(0, 150, by=10)
Y = 2*X^2 + rnorm(length(X), 0, 50^2)
xyplot(Y~X,  type=c("p", "smooth"))
```



Since our focus is on linear regression, we will need to develop appropriate methodology for capturing underlying non-linear relationships:

- Transformations (ch. 3)
- Polynomial regression (ch. 8)
- Generalized linear models (ch. 14)

1.2: Regression Models and Their Uses

A regression model is a formal means of expressing the two essential ingredients of a statistical relation:

1. A tendency of the response variable Y to vary with the predictor variable X in a systematic fashion.
2. A scattering of points around the curve of statistical relationship.

Regression and Causality

The existence of a statistical relation between the response variable Y and the explanatory (or predictor) variable X does NOT imply in any way that Y depends causally on X.

1.4: Data for Regression Analysis

Data for regression analysis may be obtained from non-experimental or experimental studies.

Observational Data

- Non-experimental
- Do not control the explanatory variable of interest
- Causal inference is hard

Experimental Data

- E.g., a completely randomized design
- Researcher controls the treatment assignment
- Causal inference is much easier to justify
    - Treatment groups are identical aside from the treatment assignment; if one group does better, it is *because of* the treatment.

Use of Computers

Computers make statistics much less tedious; use a computer.

Your book gives output and graphics from BMDP, MINITAB, SAS, SPSS, SYSTAT, JMP, S-Plus, and MATLAB.

I will give code, output, and graphics from `R`, which is basically the same as `S-Plus` except it is free!

Data Set for Warm Up Questions: SHS

The Canadian Survey of Household Spending is carried out annually across Canada.

(http://dli-idd-nesstar.statcan.gc.ca.proxy.library.upei.ca/webview/)

The main purpose of the survey is to obtain detailed information about household spending. Information is also collected about dwelling characteristics as well as household equipment.

The survey data are used by the following groups:

- Government departments use the data to help formulate policy;
- Community groups, social agencies and consumer groups use the data to support their positions and to lobby governments for social changes;
- Lawyers and their clients use the data to determine what is fair for child support and other compensation;
- Labour and contract negotiators rely on the data when discussing wage and cost-of-living clauses;
- Individuals and families can use the data to compare their spending habits with those of similar types of households.

```
#A subset of the latest Survey of Household Spending data are displayed
spending_subset %>% datatable()
```

Show 20 entries Search:

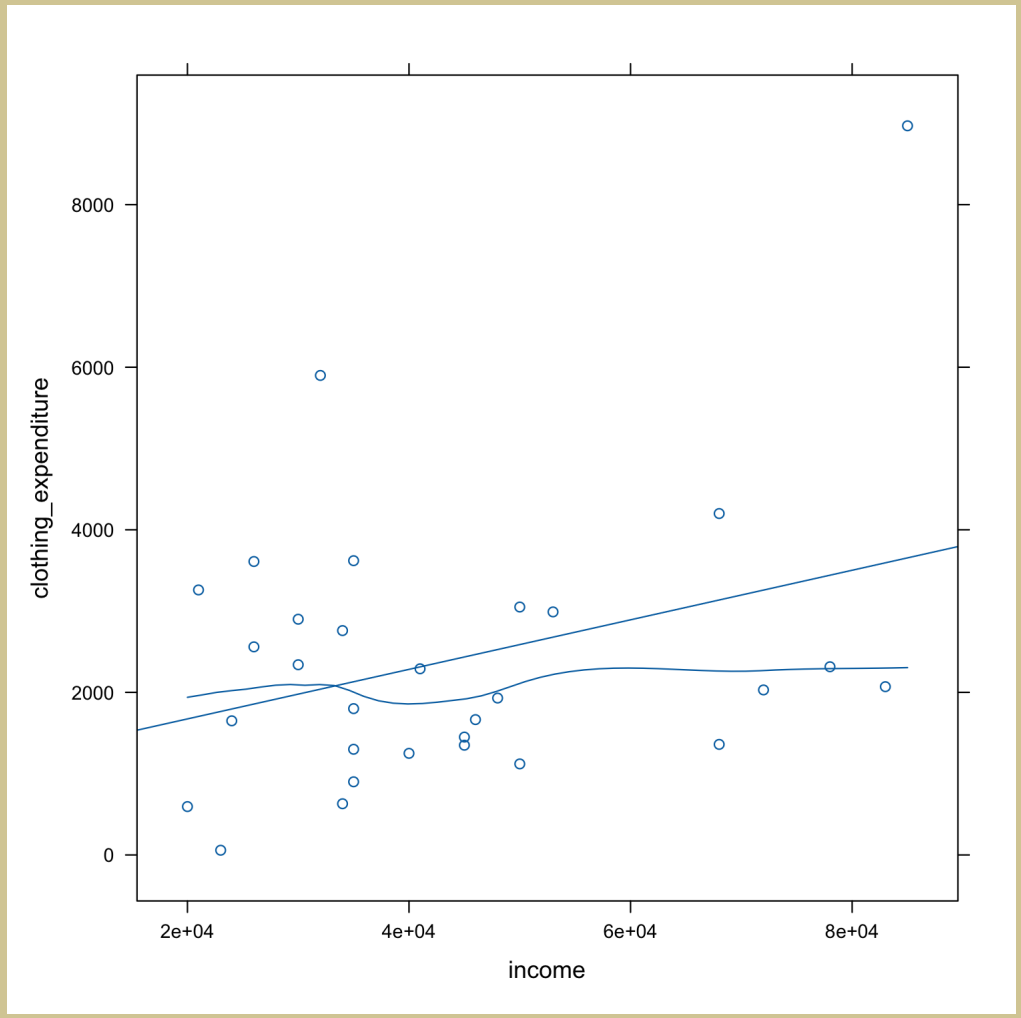
	province	type_of_dwelling	income	marital_status	age_group
1	NL	single_detached	68000	never_married	30-34
2	NL	single_detached	48000	never_married	25-29
3	NL	single_detached	30000	married	35-39
4	NL	row_house	30000	never_married	30-34
5	NL	single_detached	35000	married	25-29
6	NL	single_detached	26000	married	25-29
7	NL	single_detached	26000	other	55-59

Showing 1 to 20 of 30 entries Previous 1 2 Next

We are interested in the potential relationship between the income of working Canadians and the amount that they spend on clothing in a year.

Income and Clothing Expenditure for a small subset of the Survey of Household Spending are displayed below:

```
xyplot(clothing_expenditure~income, data=spending_subset,  type=c("p",
```



Data Set C.2: CDI

This data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States.

Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county.

Counties with missing data were deleted from the data set.

```
cdi %>% datatable()
```

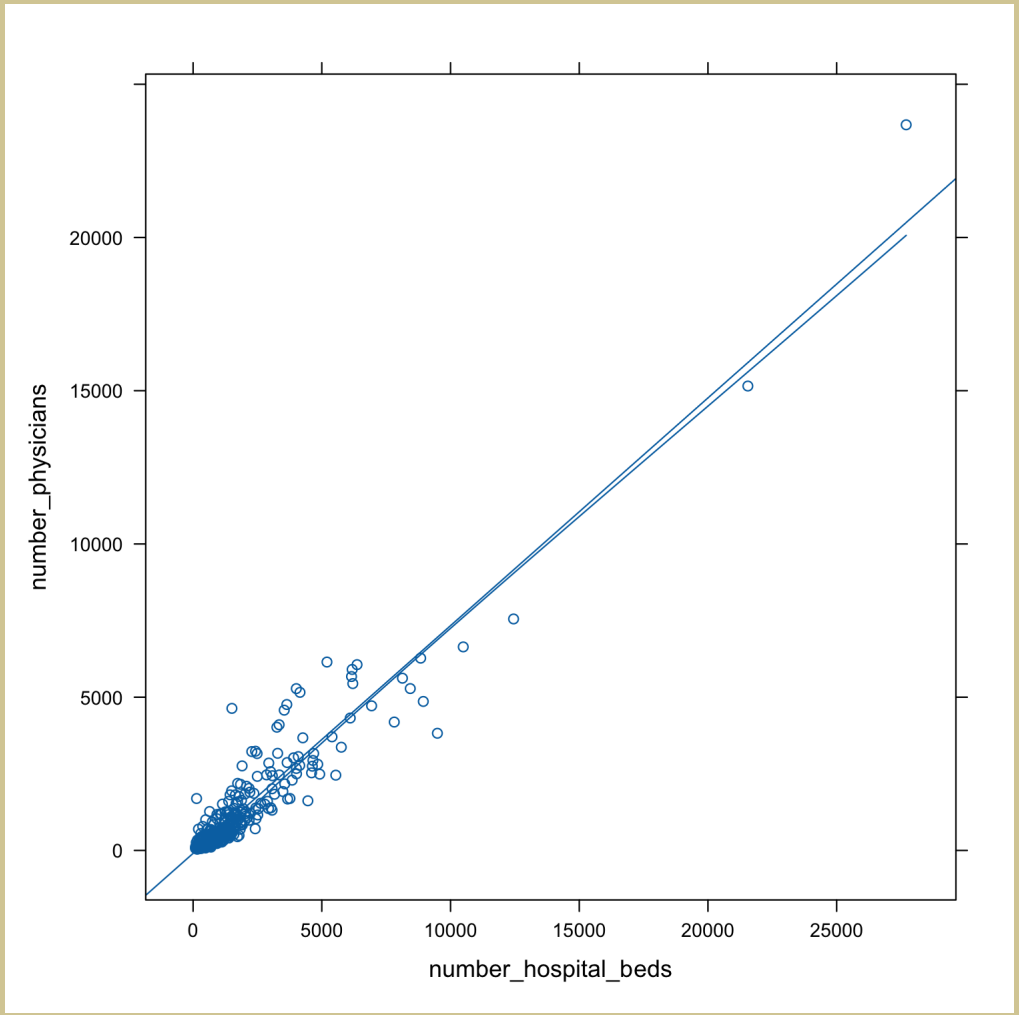
Show 20 entries Search:

	county	state	land_area	population	pop_18_to_34	pop_65	r
1	Los_Angeles	CA	4060	8863164		32.1	
2	Cook	IL	946	5105067		29.2	
3	Harris	TX	1729	2818199		31.3	
4	San_Diego	CA	4205	2498016		33.5	
5	Orange	CA	790	2410556		32.6	
6	Kings	NY	71	2300664		28.3	
7	Maricopa	AZ	9204	2122101		29.2	

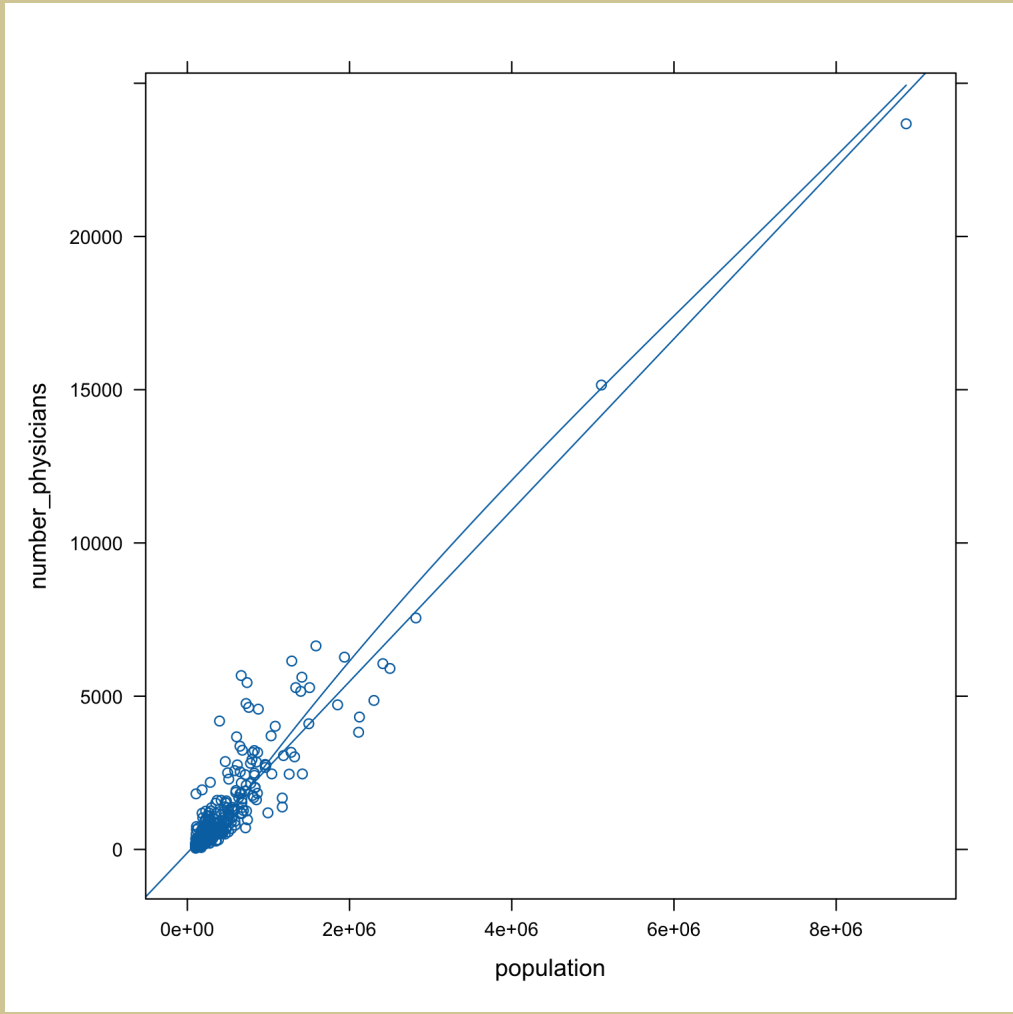
Showing 1 to 20 of 440 entries

Previous 1 2 3 4 5 … 22 Next

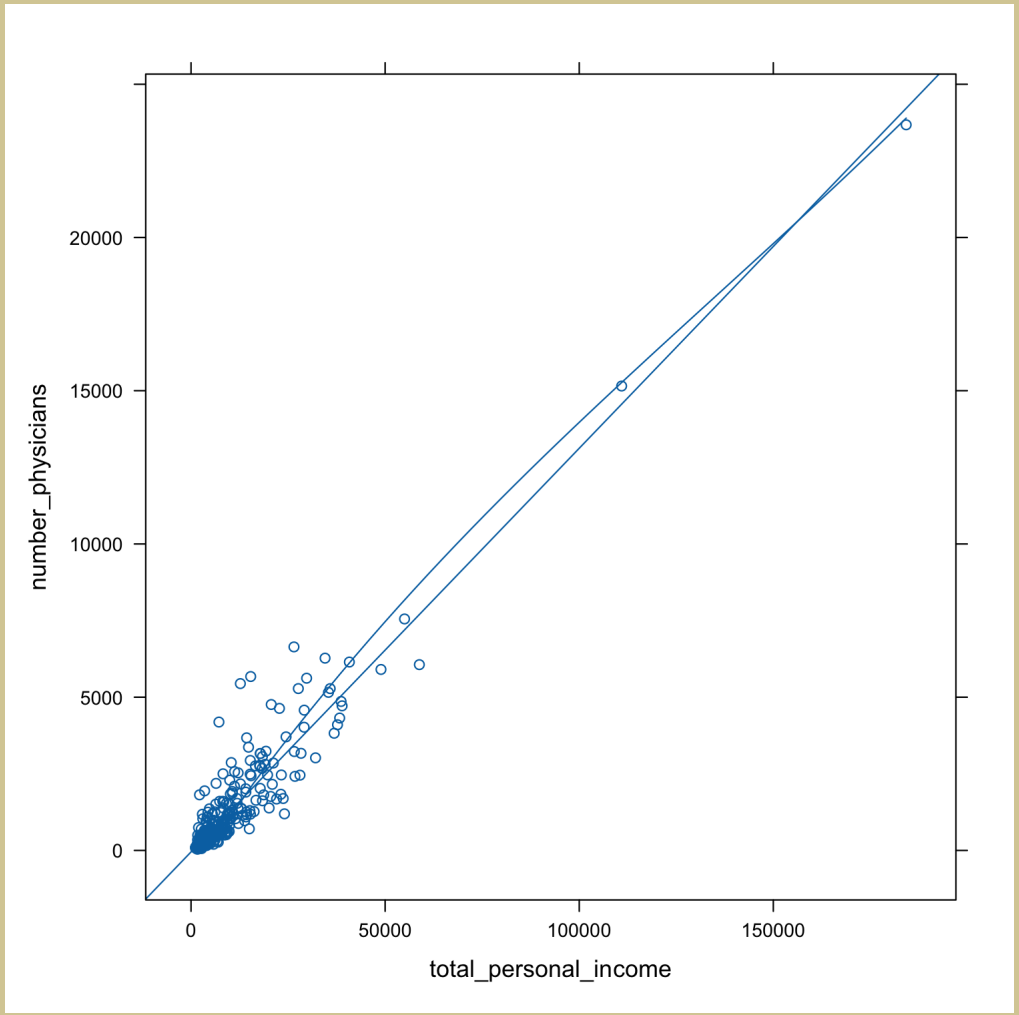
```
xyplot(number_physicians ~ number_hospital_beds, data=cdi,  type=c("p",
```



```
xyplot(number_physicians ~ population, data=cdi,  type=c("p", "r", "smo
```



```
xyplot(number_physicians ~ total_personal_income, data=cdi,  type=c("p"
```



Recap: Sections 1.1, 1.2, 1.4

After Sections 1.1, 1.2, 1.4, you should be able to

- Describe the uses of regression analysis
- Contrast regression vs causation
- Identify observational and experimental data and contrast these with respect to causation

Learning Objectives for Sections 1.3, 1.5, 1.6

After Sections 1.3, 1.5, 1.6, you should be able to

- Label and interpret the components of a regression model
- Apply the method of least squares
- Define point estimates of mean response and residuals

1.3: Simple Linear Regression Model with Distribution of Error Terms Unspecified

Want to find parameters for a function of the form

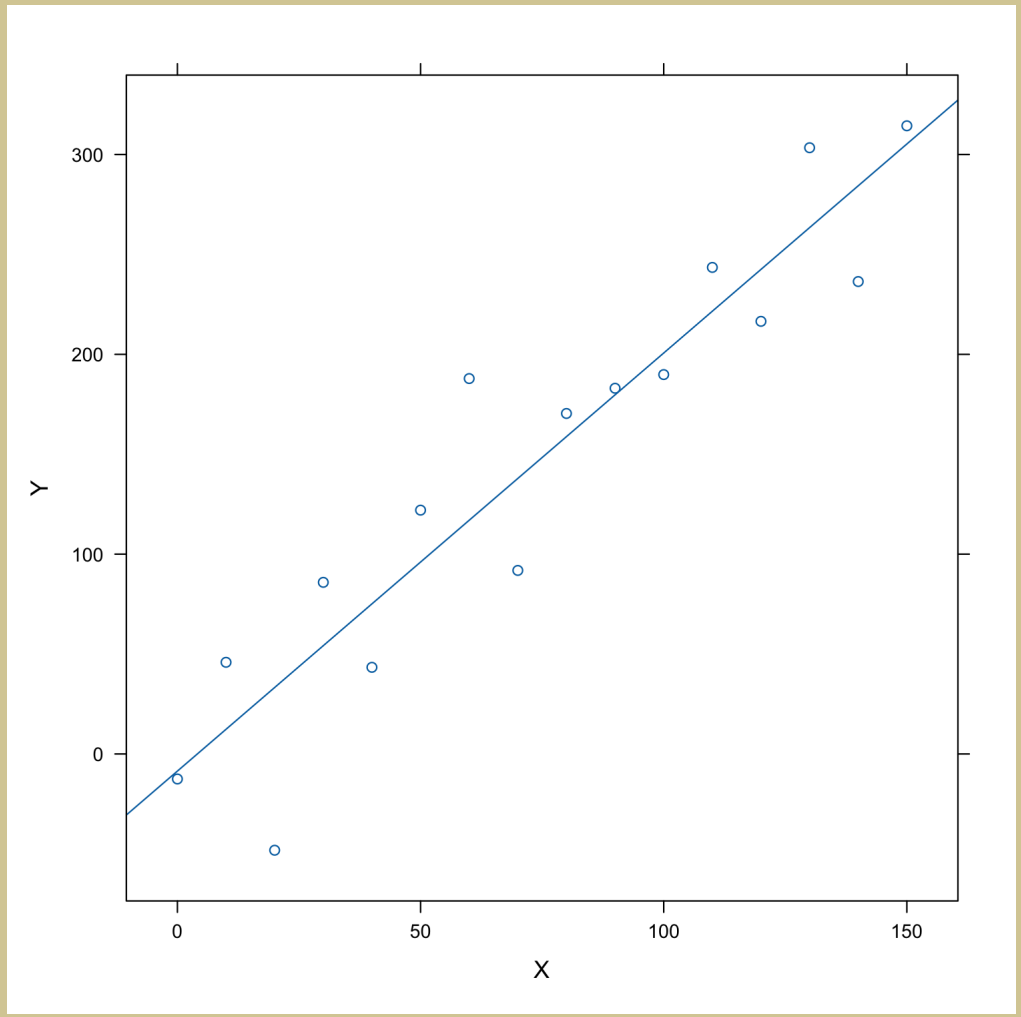
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where the distribution of error random variable not specified.

- Y_i value of the response variable in the ith trial
- β_0 and β_1 are parameters
- X_i is a known constant, the value of the predictor variable in the ith trial
- ε_i is a random error term with mean $E(\varepsilon_i) = 0$ and variance $Var(\varepsilon_i) = \sigma^2$
- ε_i and ε_j are uncorrelated
- $i = 1, \ldots, n$

This is Simple and Linear

```
X = seq(0, 150, by=10)
Y = 2*X + rnorm(length(X), 0, 50)
xyplot(Y~X,  type=c("p", "r"))
```



Important Features of Model

- The response Y_i is the sum of two components:
  - Constant term $\beta_0 + \beta_1 X_i$
  - Random term ε_i

Y_i is, therefore, a random variable

- The expected response is

$$E[Y_i] = E[\beta_0 + \beta_1 X_i + \varepsilon_i] = \beta_0 + \beta_1 X_i + E[\varepsilon_i] = \beta_0 + \beta_1 X_i$$

  - Thus, when the level of X in the ith trial is X_i, Y_i comes from a probability distribution whose mean is

$$E[Y_i] = \beta_0 + \beta_1 X_i$$

  - The regression function relates the means of the probability distributions of Y for a given X to the level of X, so our model is therefore

$$E[Y] = \beta_0 + \beta_1 X$$

- The response Y_i in the ith trial exceeds or falls short of the value of the regression function by the error term amount ε_i.

- The error terms ε_i are assumed to have a constant variance σ^2.
  - It therefore follows that the responses Y_i have the same constant variance:

$$\sigma^2\{Y_i\} = \sigma^2\{\beta_0 + \beta_1 X_i + \varepsilon_i\} = \sigma^2\{\varepsilon_i\} = \sigma^2$$

  - Thus, our regression model assumes that the probability distributions of Y have the same variance σ^2 regardless of the predictor variable X.

- The error terms are assumed to be uncorrelated. Since the error terms ε_i and ε_j are uncorrelated, so are the responses Y_i and Y_j.

In summary, this regression model implies that the responses Y_i come from probability distributions

- whose means are $E[Y_i] = \beta_0 + \beta_1 X_i$ and
- whose variances are σ^2, the same for all levels of X.

Further, any two responses Y_i and Y_j are uncorrelated.

Meaning of Regression Parameters

The parameters β_0 and β_1 are called the regression coefficients.

β_1 is the slope of the regression line

- It indicates the change in the mean of the probability distribution of Y per unit increase in X

β_0 is the Y-intercept of the regression line

- It indicates the mean of the probability distribution of Y at $X = 0$.
- If the scope of the model does not cover $X = 0$, β_0 is not meaningful.

SHS: Meaning of Regression Parameters

The Canadian Survey of Household Spending is carried out annually across Canada.

(http://dli-idd-nesstar.statcan.gc.ca.proxy.library.upei.ca/webview/)

The main purpose of the survey is to obtain detailed information about household spending. Information is also collected about dwelling characteristics as well as household equipment.

The survey data are used by the following groups:

- Government departments use the data to help formulate policy;
- Community groups, social agencies and consumer groups use the data to support their positions and to lobby governments for social changes;
- Lawyers and their clients use the data to determine what is fair for child support and other compensation;
- Labour and contract negotiators rely on the data when discussing wage and cost-of-living clauses;
- Individuals and families can use the data to compare their spending habits with those of similar types of households.

```
#A subset of the latest Survey of Household Spending data are displayed
spending_subset %>% datatable()
```

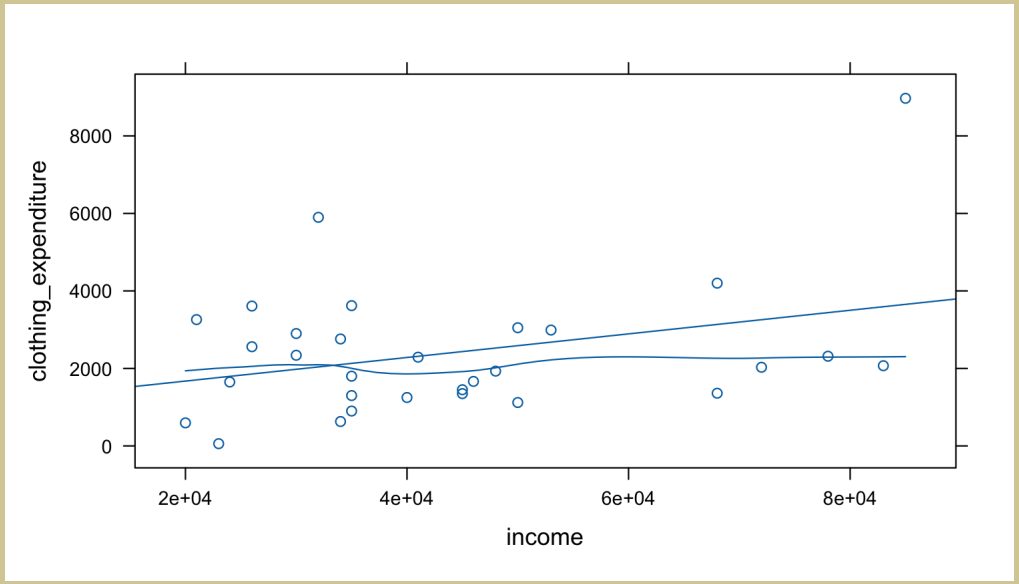
Show 20 entries Search:

	province	type_of_dwelling	income	marital_status	age_group
1	NL	single_detached	68000	never_married	30-34
2	NL	single_detached	48000	never_married	25-29
3	NL	single_detached	30000	married	35-39
4	NL	row_house	30000	never_married	30-34
5	NL	single_detached	35000	married	25-29
6	NL	single_detached	26000	married	25-29
7	NL	single_detached	26000	other	55-59

Showing 1 to 20 of 30 entries Previous 1 2 Next

We are interested in the potential relationship between the income of working Canadians and the amount that they spend on clothing in a year.

```
xyplot(clothing_expenditure~income, data=spending_subset,  type=c("p",
```



A preliminary regression analysis follows:

```
clothing_model = lm(clothing_expenditure~income, data=spending_subset)
clothing_model$coef
```

```
##  (Intercept)       income
## 1.063687e+03 3.049648e-02
```

How can we interpret β_1 in this setting? Be as specific as possible.

How can we interpret β_0 in this setting? Be as specific as possible.

Data Set C.2: CDI

This data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States.

Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county.

Counties with missing data were deleted from the data set.

```
cdi %>% datatable()
```

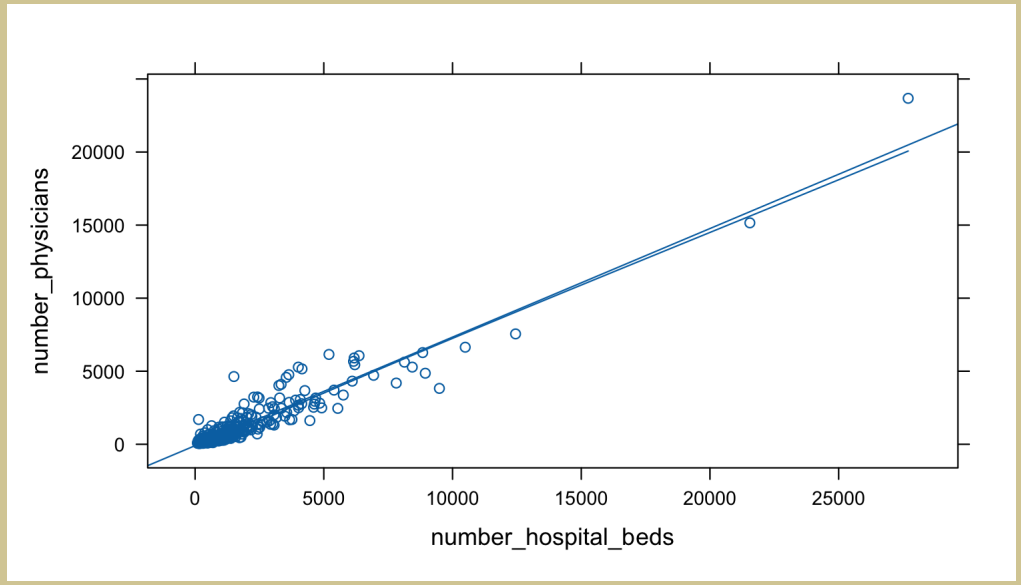
Show 20 entries Search:

	county	state	land_area	population	pop_18_to_34	pop_65	r
1	Los_Angeles	CA	4060	8863164		32.1	
2	Cook	IL	946	5105067		29.2	
3	Harris	TX	1729	2818199		31.3	
4	San_Diego	CA	4205	2498016		33.5	
5	Orange	CA	790	2410556		32.6	
6	Kings	NY	71	2300664		28.3	
7	Maricopa	AZ	9204	2122101		29.2	

Showing 1 to 20 of 440 entries

Previous 1 2 3 4 5 … 22 Next

```
mod_physician_beds = lm(number_physicians ~ number_hospital_beds, data=
xyplot(number_physicians ~ number_hospital_beds, data=cdi,  type=c("p",
```



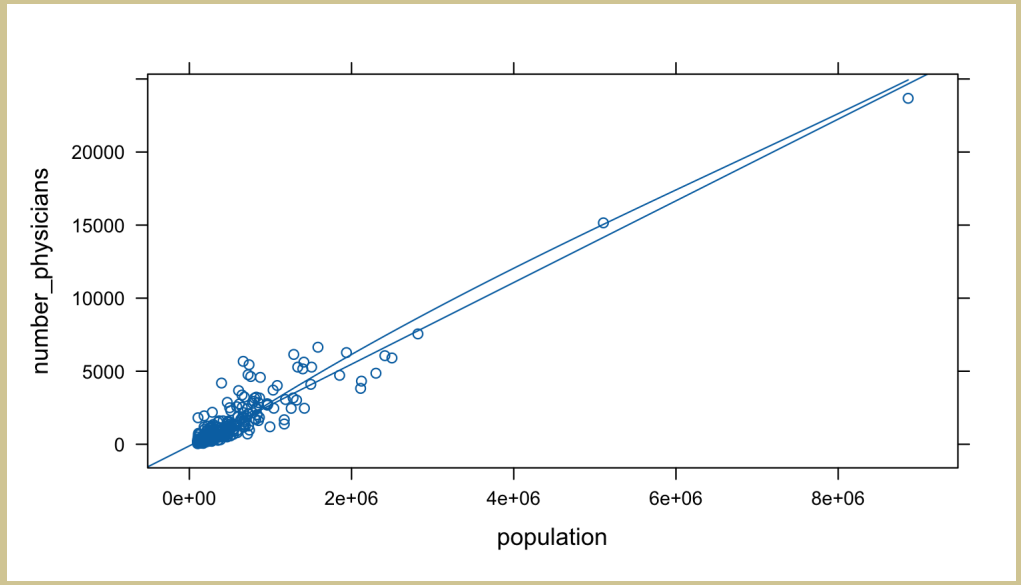

```
mod_physician_beds$coef
```

```
##          (Intercept) number_hospital_beds
##          -95.9321847            0.7431164
```

How can we interpret β_1 in this setting? Be as specific as possible.

How can we interpret β_0 in this setting? Be as specific as possible.

```
mod_physician_pop = lm(number_physicians ~ population, data=cdi)
xyplot(number_physicians ~ population, data=cdi,  type=c("p", "r", "smo
```



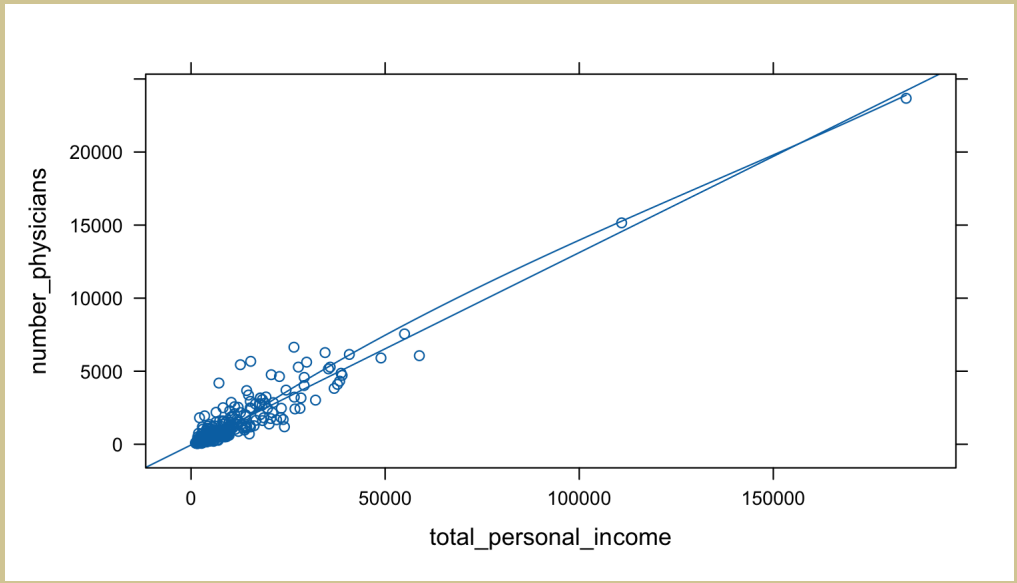

```
mod_physician_pop$coef
```

```
##   (Intercept)    population
## -1.106348e+02  2.795425e-03
```

How can we interpret β_1 in this setting? Be as specific as possible.

How can we interpret β_0 in this setting? Be as specific as possible.

```
mod_physician_income = lm(number_physicians ~ total_personal_income, da
xyplot(number_physicians ~ total_personal_income, data=cdi,  type=c("p"
```




```
mod_physician_income$coef
```

```
##           (Intercept) total_personal_income
##           -48.3948489             0.1317012
```

How can we interpret β_1 in this setting? Be as specific as possible.

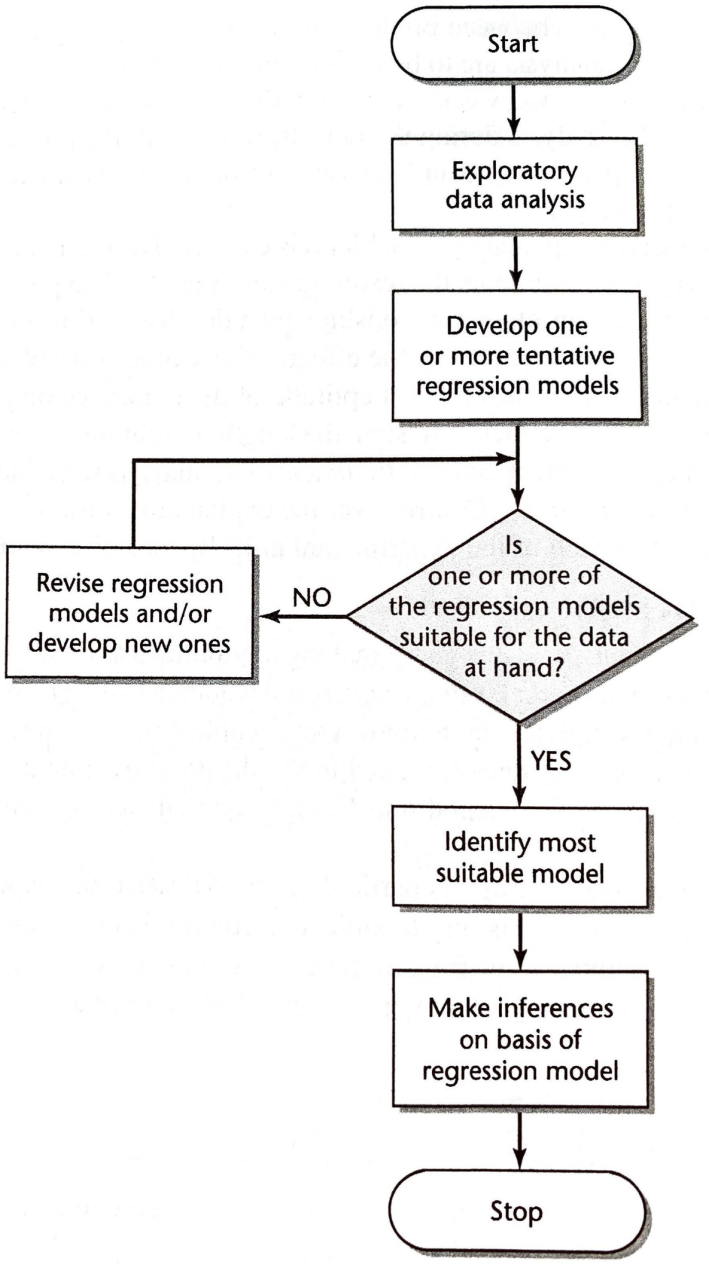
How can we interpret β_0 in this setting? Be as specific as possible.

1.5: Overview of Steps in Regression Analysis

We will begin by discussing inferences based on a regression model that is considered to be appropriate.

However, throughout the course, we will be stepping back from this and considering the development of a suitable regression model.

FIGURE 1.8
Typical Strategy for Regression Analysis.



1.6: Estimation of Regression Function

Assuming we know the form of the regression function,

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

how do we go about estimating the parameters?

Example:

An experimenter gave three subjects a very difficult task. Data on the age of the subject (X) and on the number of attempts to accomplish the task before giving up (Y) follow:

Subject i	1	2	3
Age X_i	20	55	30
Number of Attempts Y_i	5	12	10

Note:

- $n = 3$
- $(X_1, Y_1) = (20, 5)$
- $(X_2, Y_2) = (55, 12)$
- $(X_3, Y_3) = (30, 10)$

Method of Least Squares

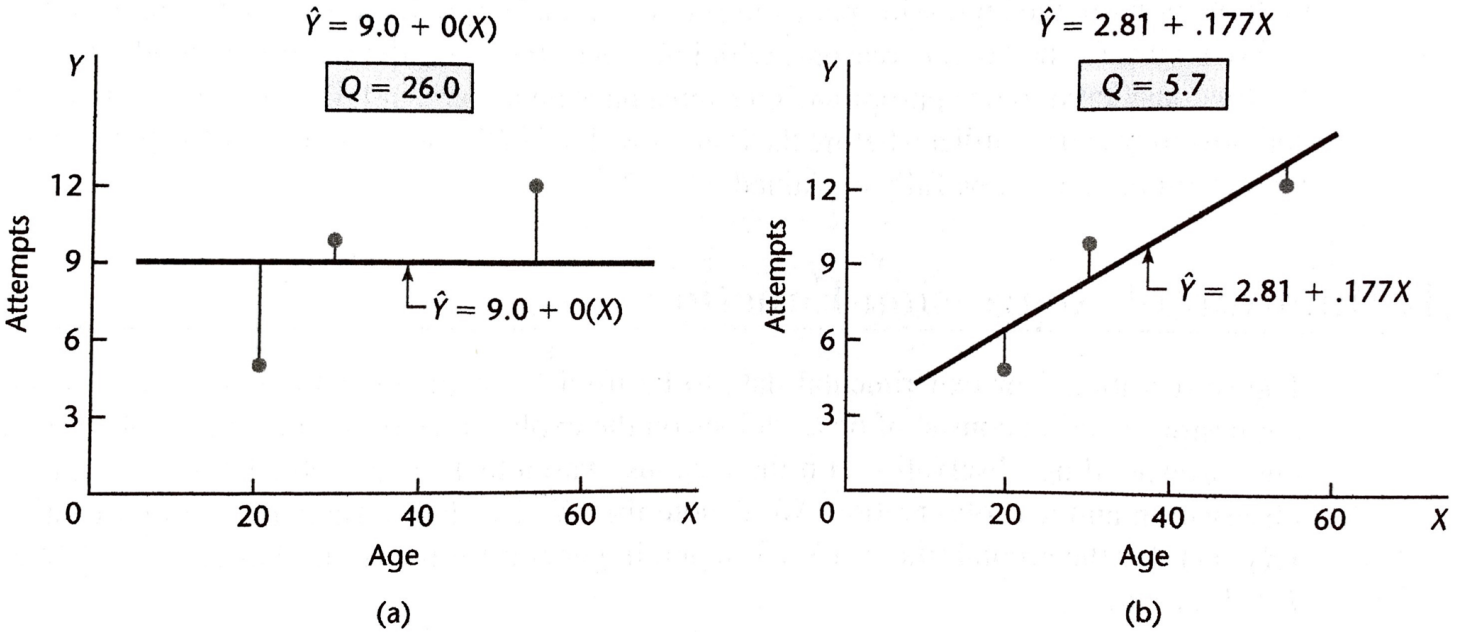
Goal: estimate β_0 and β_1 with b_0 and b_1 if they result in small *deviations*:

$$Y_i - (b_0 + b_1 X_i).$$

- That is, we will make Y_i and $b_0 + b_1 X_i$ close for all i for our observations (X_i, Y_i).
- In particular, the *method of least squares* seeks to minimize the sum of the n squared deviations:

$$Q = \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2.$$

FIGURE 1.9 **Illustration of Least Squares Criterion Q for Fit of a Regression Line—Persistence Study Example.**



Finding Least Squares Point Estimates

We want b_0 and b_1, the values of β_0 and β_1 that minimize

$$Q = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2.$$

In nice situations, this can be accomplished by taking the derivative of the function and finding where it equals 0:

$$0 = \left.\frac{\partial Q}{\partial \beta_0}\right|_{\beta=b} = -2 \sum [Y_i - (b_0 + b_1 X_i)]$$

$$0 = \left.\frac{\partial Q}{\partial \beta_1}\right|_{\beta=b} = -2 \sum X_i [Y_i - (b_0 + b_1 X_i)]$$

After some simplification, we get the so-called *normal equations*:

$$\sum Y_i = n b_0 + b_1 \sum X_i$$

$$\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2$$

These *normal equations* can be solved simultaneously for b_0 and b_1:

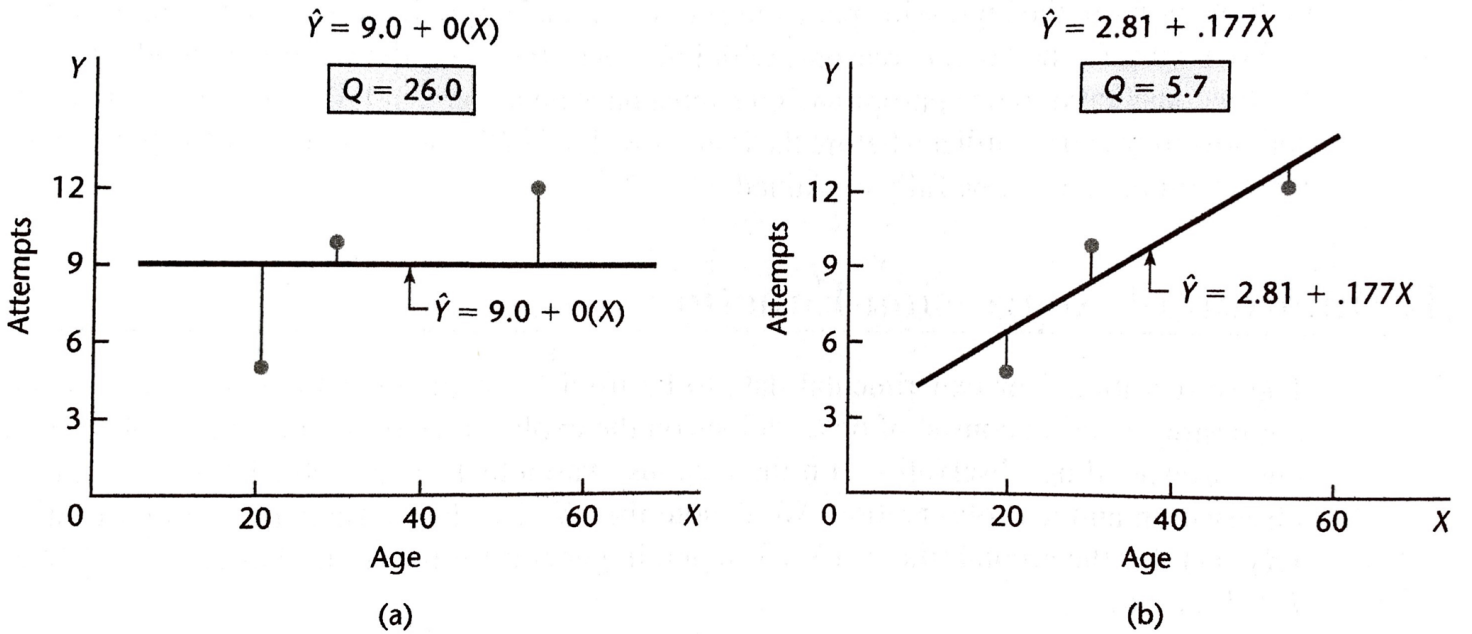
$$b_1 = \frac{\sum(X_i - \bar{X}_i)\sum(Y_i - \bar{Y}_i)}{\sum(X_i - \bar{X}_i)^2}$$
$$b_0 = \bar{Y} - b_1\bar{X}$$

where

$$\bar{Y} = \frac{1}{n}\sum Y_i$$
$$\bar{X} = \frac{1}{n}\sum X_i$$

Properties of Least Squares Estimators

FIGURE 1.9 Illustration of Least Squares Criterion Q for Fit of a Regression Line—Persistence Study Example.



- The least squares estimators b_0 and b_1 are unbiased:

$$E[b_0] = \beta_0 \qquad \text{and} \qquad E[b_1] = \beta_1.$$

- b_0 and b_1 have minimum variance among all unbiased linear estimators

Point Estimation of Mean Response

Given sample estimators b_0 and b_1 of the parameters in the regression function

$$E[Y] = \beta_0 + \beta_1 X,$$

we estimate the regression function as follows

$$\boxed{\hat{Y} = b_0 + b_1 X,}$$

where $\hat{Y}$ is the value of the estimated regression function at the level X of the predictor variable.

$\hat{Y}$ is an unbiased estimator of $E[Y]$ with minimum variance in the class of unbiased linear estimators.

- we call the *value* of the response variable a response
- we call $E[Y]$ the mean response

For observed cases in the study, we will call $\hat{Y}_i = b_0 + b_1 X_i$ the fitted value for the ith case ($i = 1, \ldots, n$)

This *fitted value* is different than the *observed value* Y_i

The difference between the observed value and the fitted value is the ith residual:

$$\boxed{e_i = Y_i - \hat{Y}_i} = Y_i - (b_0 + b_1 X_i)$$

Note that the *model error term* $\varepsilon_i = Y_i - E[Y_i]$ is not the same as the *residual* $e_i = Y_i - \hat{Y}_i$

SHS: least squares estimates, estimates of mean response, observed values, fitted values, residuals

```
#A subset of the latest Survey of Household Spending data are displayed
spending_subset %>% datatable(options=list(scrollY=200))
```

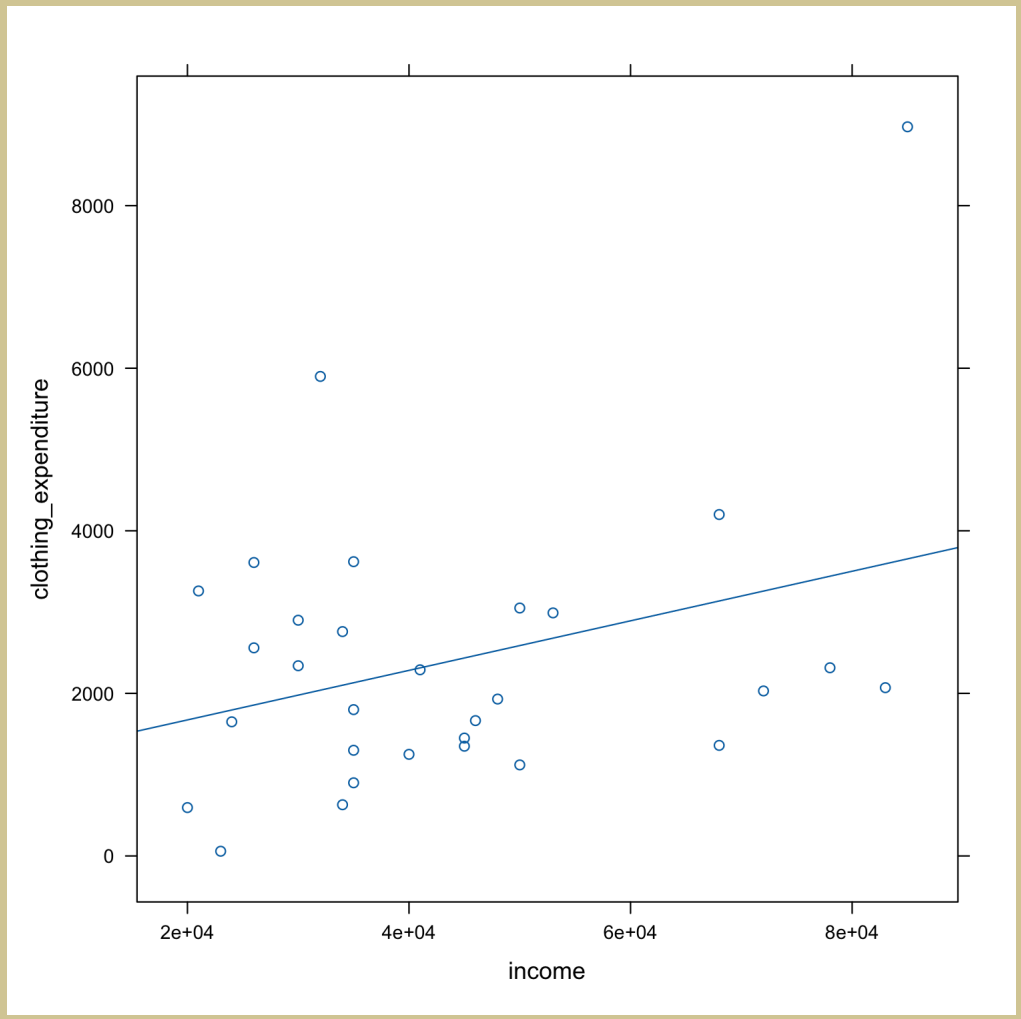
Show 20 entries Search:

	province	type_of_dwelling	income	marital_status	age_group
1	NL	single_detached	68000	never_married	30-34
2	NL	single_detached	48000	never_married	25-29
3	NL	single_detached	30000	married	35-39
4	NL	row_house	30000	never_married	30-34
5	NL	single_detached	35000	married	25-29

Showing 1 to 20 of 30 entries Previous 1 2 Next

Income and Clothing Expenditure for a small subset of the Survey of Household Spending are displayed below:

```
xyplot(clothing_expenditure~income, data=spending_subset,  type=c("p",
```



A preliminary regression analysis follows:

```
clothing_model = lm(clothing_expenditure~income, data=spending_subset)
clothing_model$coef
```

```
##  (Intercept)       income
## 1.063687e+03 3.049648e-02
```

```
tibble(X=spending_subset$income, Y = spending_subset$clothing_expenditu
```

Show 20 entries Search:

	X	Y	Yhat	e
1	68000	4200	3137.44787	1062.55213
2	48000	1930	2527.51831	-597.51831
3	30000	2340	1978.5817	361.4183
4	30000	2900	1978.5817	921.4183
5	35000	1300	2131.06409	-831.06409
6	26000	3610	1856.59579	1753.40421
7	26000	2560	1856.59579	703.40421

Showing 1 to 20 of 30 entries Previous 1 2 Next

CDI: least squares estimates, estimates of mean response, observed values, fitted values, residuals

This data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States.

Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county.

Counties with missing data were deleted from the data set.

```
cdi %>% datatable()
```

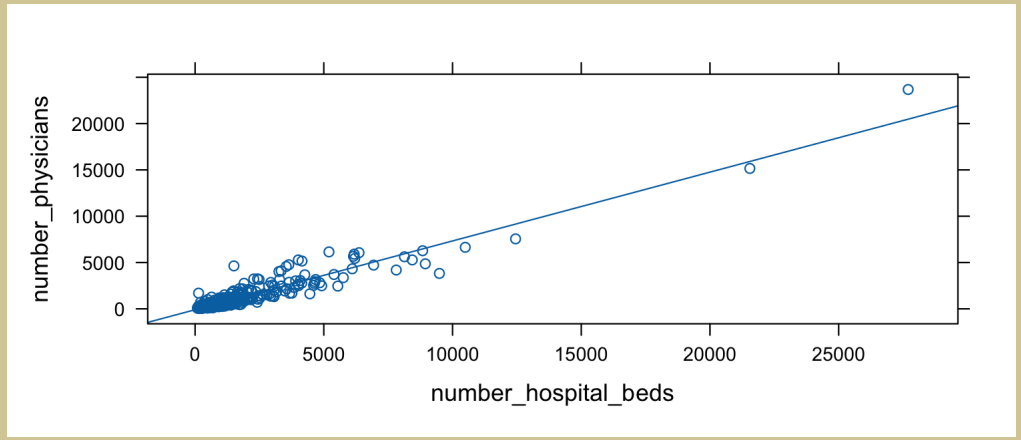
Show 20 entries Search:

	county	state	land_area	population	pop_18_to_34	pop_65	r
1	Los_Angeles	CA	4060	8863164		32.1	
2	Cook	IL	946	5105067		29.2	
3	Harris	TX	1729	2818199		31.3	
4	San_Diego	CA	4205	2498016		33.5	
5	Orange	CA	790	2410556		32.6	
6	Kings	NY	71	2300664		28.3	
7	Maricopa	AZ	9204	2122101		29.2	

Showing 1 to 20 of 440 entries

Previous 1 2 3 4 5 … 22 Next

```
mod_physician_beds = lm(number_physicians ~ number_hospital_beds, data=
xyplot(number_physicians ~ number_hospital_beds, data=cdi,  type=c("p",
```




```
mod_physician_beds$coef
```

```
##          (Intercept) number_hospital_beds
##          -95.9321847            0.7431164
```

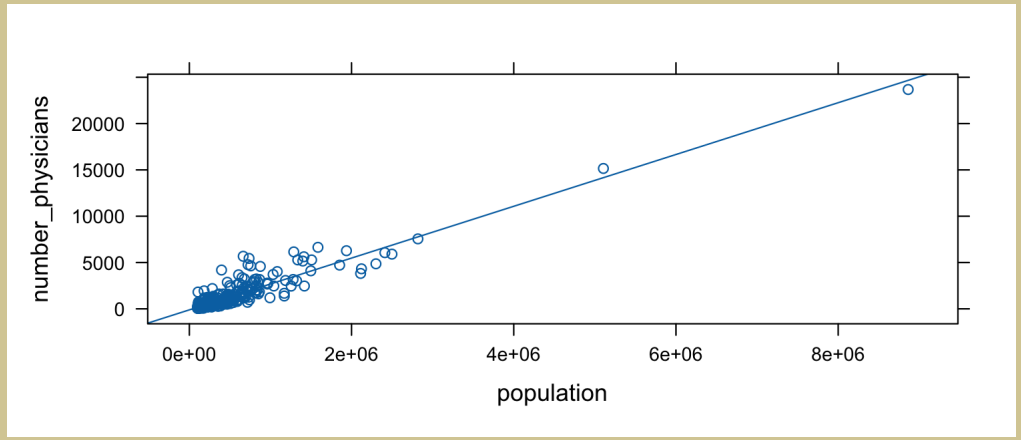
```
tibble(X1=cdi$number_hospital_beds, Y = cdi$number_physicians, Yhat1 =
```

Show 20 entries Search:

	X1	Y	Yhat1	e1
1	27700	23677	20488.39331	3188.60669

Showing 1 to 20 of 440 entries

```
mod_physician_pop = lm(number_physicians ~ population, data=cdi)
xyplot(number_physicians ~ population, data=cdi,  type=c("p", "r"))
```




```
mod_physician_pop$coef
```

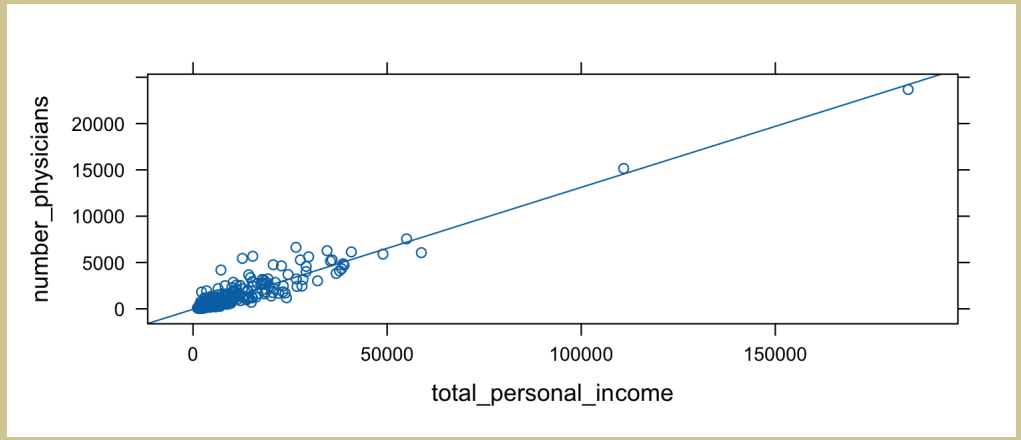
```
##   (Intercept)    population
## -1.106348e+02  2.795425e-03
```

```
tibble(X2=cdi$population, Y = cdi$number_physicians, Yhat2 = predict(mo
```

Show 20 entries Search:

	X2	Y	Yhat2	e2
1	8863164	23677	24665.67437	-988.67437

```
mod_physician_income = lm(number_physicians ~ total_personal_income, da
xyplot(number_physicians ~ total_personal_income, data=cdi,  type=c("p"
```




```
mod_physician_income$coef
```

```
##           (Intercept) total_personal_income
##           -48.3948489             0.1317012
```

```
tibble(X3=cdi$total_personal_income, Y = cdi$number_physicians, Yhat3 =
```

Show 20 entries Search:

	X3	Y	Yhat3	e3
1	184230	23677	24214.91523	-537.91523

Properties of Fitted Regression Line

- The regression line always goes through the point $(\bar{X}, \bar{Y})$.
- The sum of the squared residuals, $\sum_{i=1}^{n} e_i^2$, is a minimum

- The sum of the residuals is zero:

$$\sum_{i=1}^{n} e_i = 0$$

- The sum of the observed values Y_i equals the sum of the fitted values $\hat{Y}_i$:

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$$

- The sum of the weighted residuals is zero when the residual in the ith trial is weighted by the level of the predictor variable in the ith trial:

$$\sum_{i=1}^{n} X_i e_i = 0$$

- The sum of the weighted residuals is zero when the residual in the ith trial is weighted by the fitted value of the response variable in the ith trial

$$\sum_{i=1}^{n} \hat{Y}_i e_i = 0$$

SHS: Properties of Fitted Regression Line

```
#A subset of the latest Survey of Household Spending data are displayed
spending_subset %>% datatable()
```

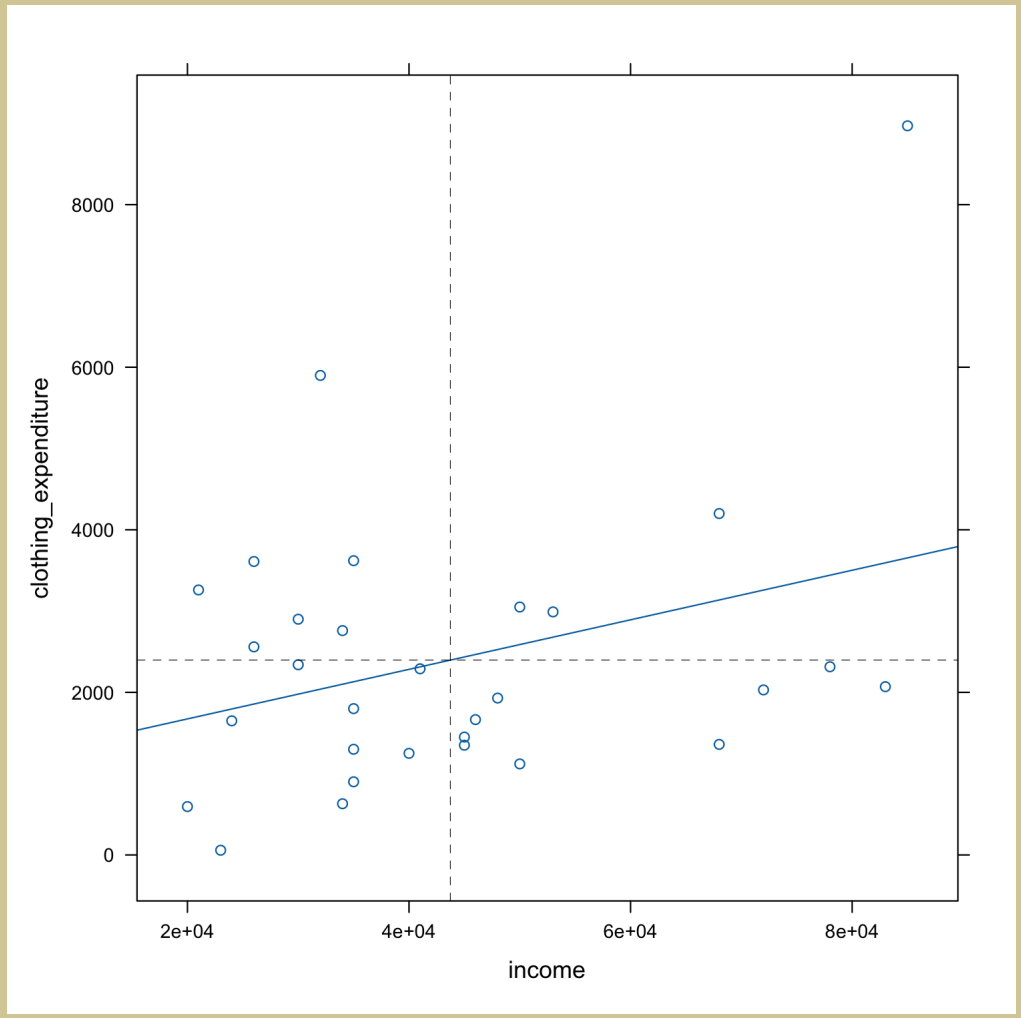
Show 20 entries Search:

	province	type_of_dwelling	income	marital_status	age_group
1	NL	single_detached	68000	never_married	30-34
2	NL	single_detached	48000	never_married	25-29
3	NL	single_detached	30000	married	35-39
4	NL	row_house	30000	never_married	30-34
5	NL	single_detached	35000	married	25-29
6	NL	single_detached	26000	married	25-29
7	NL	single_detached	26000	other	55-59

Showing 1 to 20 of 30 entries Previous 1 2 Next

```
xyplot(clothing_expenditure~income, data=spending_subset,  type=c("p",
```

```
ladd(panel.abline(h=mean(spending_subset$clothing_expenditure), v=mean(
```



```
clothing_model = lm(clothing_expenditure~income, data=spending_subset)
clothing_tble = tibble(X=spending_subset$income, Y = spending_subset$cl
clothing_tble %>% round() %>% datatable(options=list(scrollY=200))
```

Show 20 entries Search:

	X	Y	Yhat	e	Xe	YhatE
1	68000	4200	3137	1063	72253545	3333702
2	48000	1930	2528	-598	-28680879	-1510238
3	30000	2340	1979	361	10842549	715096
4	30000	2900	1979	921	27642549	1823101
5	35000	1300	2131	-831	-29087243	-1771051

Showing 1 to 20 of 30 entries Previous 1 2 Next

```
round(colSums(clothing_tble), 5)[-1]
```

```
##     Y  Yhat     e    Xe YhatE
## 71922 71922     0     0     0
```

CDI: Properties of Fitted Regression Line

This data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States.

Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county.

Counties with missing data were deleted from the data set.

```
cdi %>% datatable()
```

Show 20 entries Search:

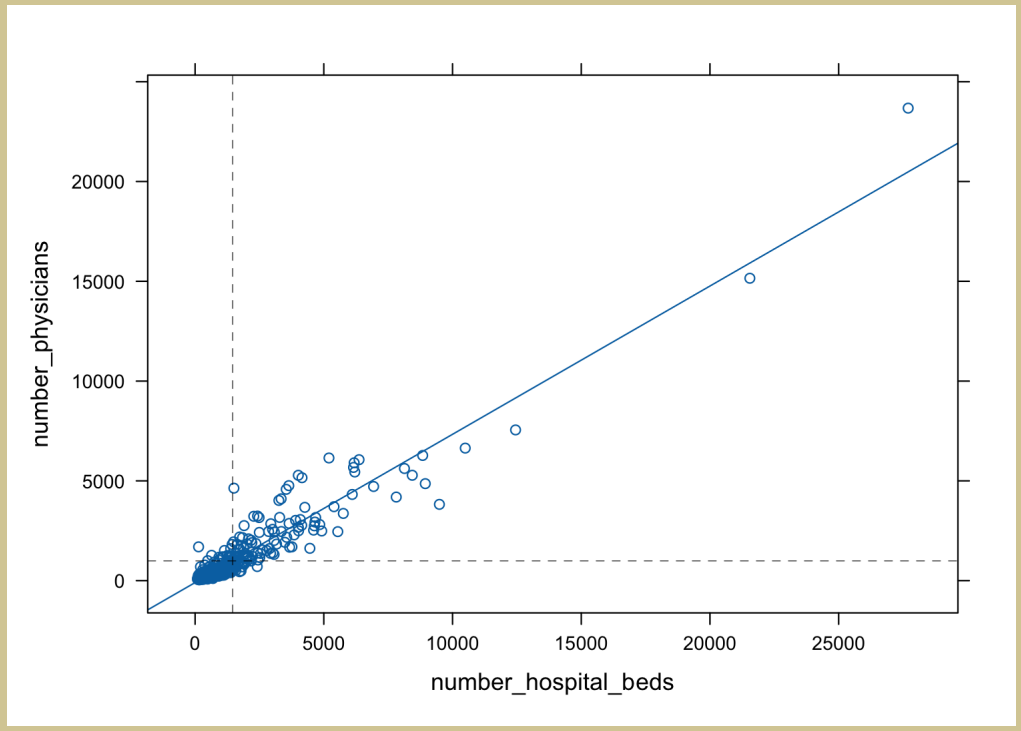
	county	state	land_area	population	pop_18_to_34	pop_65	r
1	Los_Angeles	CA	4060	8863164		32.1	
2	Cook	IL	946	5105067		29.2	
3	Harris	TX	1729	2818199		31.3	
4	San_Diego	CA	4205	2498016		33.5	
5	Orange	CA	790	2410556		32.6	
6	Kings	NY	71	2300664		28.3	
7	Maricopa	AZ	9204	2122101		29.2	

Showing 1 to 20 of 440 entries

Previous 1 2 3 4 5 … 22 Next

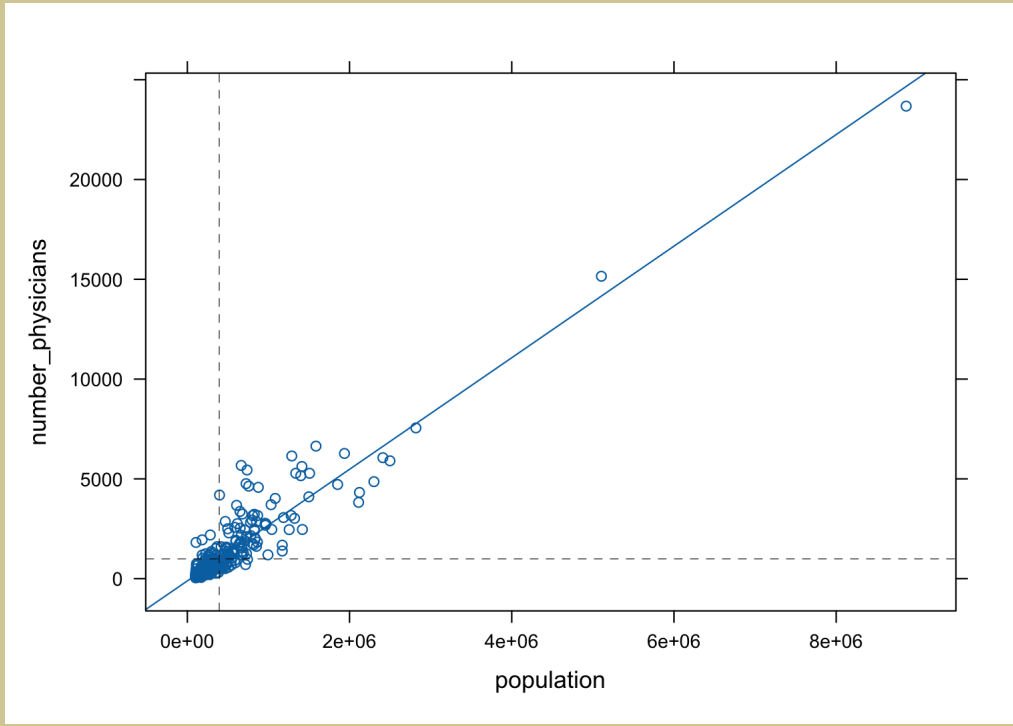
```
mod_physician_beds = lm(number_physicians ~ number_hospital_beds, data=
xyplot(number_physicians ~ number_hospital_beds, data=cdi,  type=c("p",
```

```
ladd(panel.abline(h=mean(cdi$number_physicians), v=mean(cdi$number_hosp
```



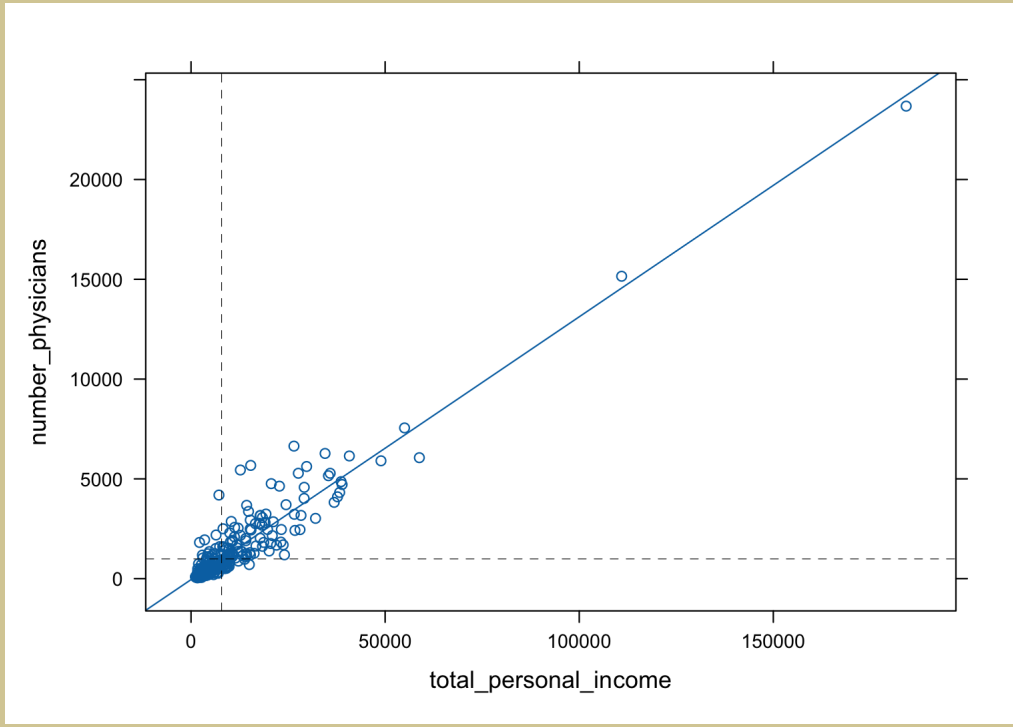
```
mod_physician_pop = lm(number_physicians ~ population, data=cdi)
xyplot(number_physicians ~ population, data=cdi,  type=c("p", "r"))
```

```
ladd(panel.abline(h=mean(cdi$number_physicians), v=mean(cdi$population)
```



```
mod_physician_income = lm(number_physicians ~ total_personal_income, da
xyplot(number_physicians ~ total_personal_income, data=cdi,  type=c("p"
```

```
ladd(panel.abline(h=mean(cdi$number_physicians), v=mean(cdi$total_perso
```



```
cdi_tble = tibble(X1=cdi$number_hospital_beds, X2=cdi$population, X3=cd
cdi_tble %>% round() %>% datatable(options=list(scrollY=150))
```

Show 20 entries Search:

	X1	X2	X3	Y	Yhat1	e1	X1e1	Yhat1E1
1	27700	8863164	184230	23677	20488	3189	88324405	6532942
2	21550	5105067	110928	15153	15918	-765	-16490646	-1218106
3	12449	2818199	55003	7553	9155	-1602	-19944847	-1466764
4	6179	2498016	48931	5905	4496	1409	8707544	633553

Showing 1 to 20 of 440 entries

Previous 1 2 3 4 5 ... 22 Next

```
round(colMeans(cdi_tble), 5)[-c(1,2,3)]
```

```
##         Y    Yhat1       e1     X1e1  Yhat1E1    Yhat2       e2     X2e2
## 987.9977 987.9977   0.0000   0.0000   0.0000 987.9977   0.0000   0.0000
##  Yhat2E2    Yhat3       e3     X3e3  Yhat3E3
##   0.0000 987.9977   0.0000   0.0000   0.0000
```

Recap: Sections 1.4-1.6

After Sections 1.4-1.6, you should be able to

- Label and interpret the components of a regression model
- Apply the method of least squares
- Define point estimates of mean response and residuals

Learning Objectives for Sections 1.7-1.8

After Sections 1.7-1.8, you should be able to

- Define the normal error regression model
- Define and interpret SSE and MSE
- Apply the method of maximum likelihood

1.7: Estimation of Error Terms Variance σ^2

The variance σ^2 of the error terms ε_i in the regression model

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

needs to be estimated in order to develop interval estimates and to perform inference.

Point Estimator of σ^2 in a Single Population

In STAT 1910, we estimated the mean, μ, and variance, σ^2, of a population by using the sample statistics $\bar{Y}$ and s^2.

The Sample Variance Estimator consists of a the sum of squared deviations divided by the degrees of freedom associated with it:

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$$

The Sum of Squares is

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

and the degrees of freedom here is $n-1$ because one degree of freedom is lost by using $\bar{Y}$ as an estimate of the unknown population mean μ.

- s^2 is called a mean square because the sum of squares has been divided by the appropriate degrees of freedom
- s^2 is an unbiased estimator of σ^2.

Point Estimator of σ^2 in a Regression Model

In a regression model, each Y_i

- comes from a different probability distribution with a different mean that depend on the level X_i
- has variance σ^2 (the same as for the error term ε_i)

Thus, the deviation of an observation Y_i must be calculated around its own estimated mean:

$$\boxed{SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2,}$$

where *SSE* stands for error sum of squares or residual sum of squares.

- The sum of squares *SSE* has $n - 2$ degrees of freedom associated with it
  - 2 degrees of freedom are lost as both β_0 and β_1 had to be estimated in order to obtain $\hat{Y}_i$
- The corresponding mean square error (or mean square residual) is

$$\boxed{MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2},}$$

- *MSE* is an unbiased estimator of σ^2:

$$E[MSE] = \sigma^2.$$

An estimator of the standard deviation is simply $s = \sqrt{MSE}$

SHS: SSE and MSE

The Canadian Survey of Household Spending is carried out annually across Canada.

(http://dli-idd-nesstar.statcan.gc.ca.proxy.library.upei.ca/webview/)

The main purpose of the survey is to obtain detailed information about household spending. Information is also collected about dwelling characteristics as well as household equipment.

The survey data are used by the following groups:

- Government departments use the data to help formulate policy;
- Community groups, social agencies and consumer groups use the data to support their positions and to lobby governments for social changes;
- Lawyers and their clients use the data to determine what is fair for child support and other compensation;
- Labour and contract negotiators rely on the data when discussing wage and cost-of-living clauses;
- Individuals and families can use the data to compare their spending habits with those of similar types of households.

```
#A subset of the latest Survey of Household Spending data are displayed
spending_subset %>% datatable()
```

Show 20 entries Search:

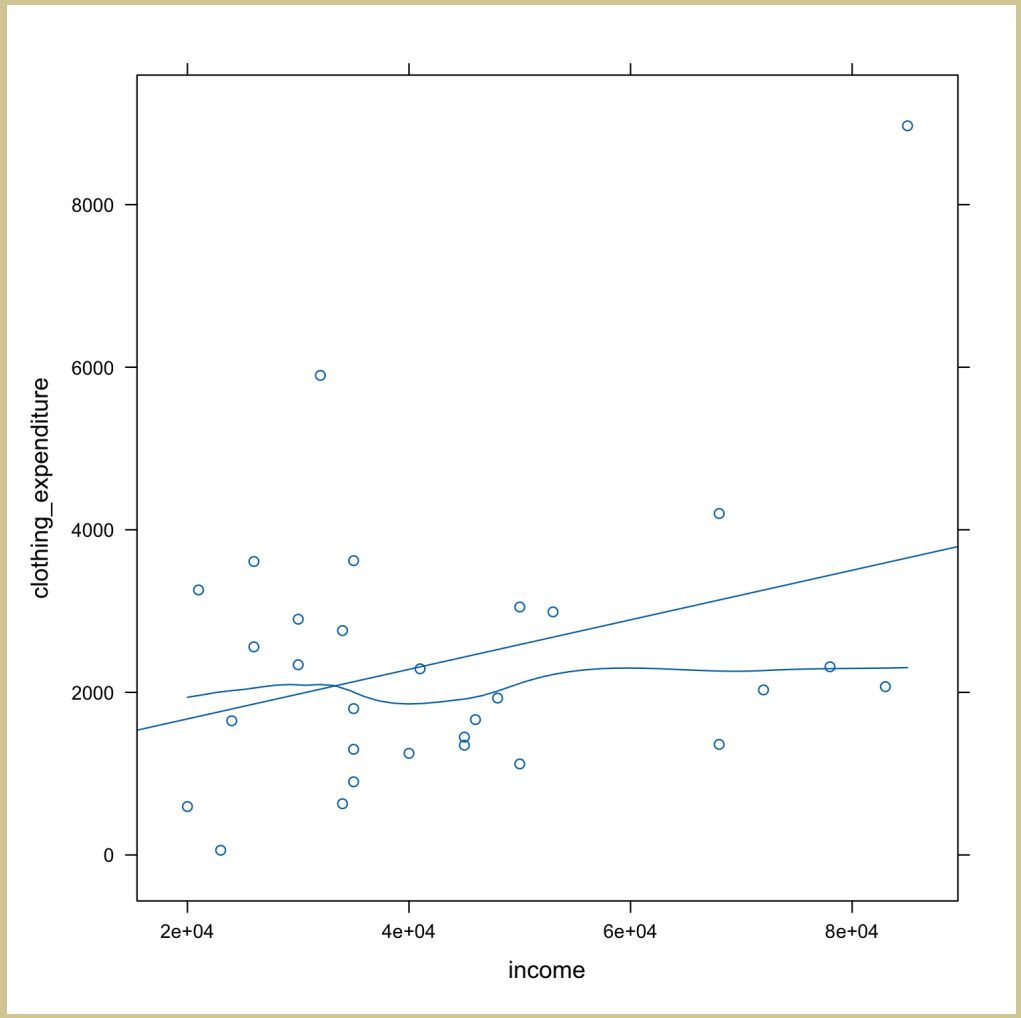
	province	type_of_dwelling	income	marital_status	age_group
1	NL	single_detached	68000	never_married	30-34
2	NL	single_detached	48000	never_married	25-29
3	NL	single_detached	30000	married	35-39
4	NL	row_house	30000	never_married	30-34
5	NL	single_detached	35000	married	25-29
6	NL	single_detached	26000	married	25-29
7	NL	single_detached	26000	other	55-59

Showing 1 to 20 of 30 entries Previous 1 2 Next

We are interested in the potential relationship between the income of working Canadians and the amount that they spend on clothing in a year.

Income and Clothing Expenditure for a small subset of the Survey of Household Spending are displayed below:

```
xyplot(clothing_expenditure~income, data=spending_subset,  type=c("p",
```



A preliminary regression analysis follows:

```
clothing_model = lm(clothing_expenditure~income, data=spending_subset)

SSE = sum(residuals(clothing_model)^2)

df = (summary(clothing_model)$df[2])

SSE /df
```

```
## [1] 2764833
```

```
s = summary(clothing_model)$sigma
s
```

```
## [1] 1662.779
```

```
s^2
```

```
## [1] 2764833
```

```
predict(clothing_model, newdata=data.frame(income=60000))
```

```
##        1
## 2893.476
```

CDI: SSE and MSE

This data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States.

Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county.

Counties with missing data were deleted from the data set.

```
cdi %>% datatable()
```

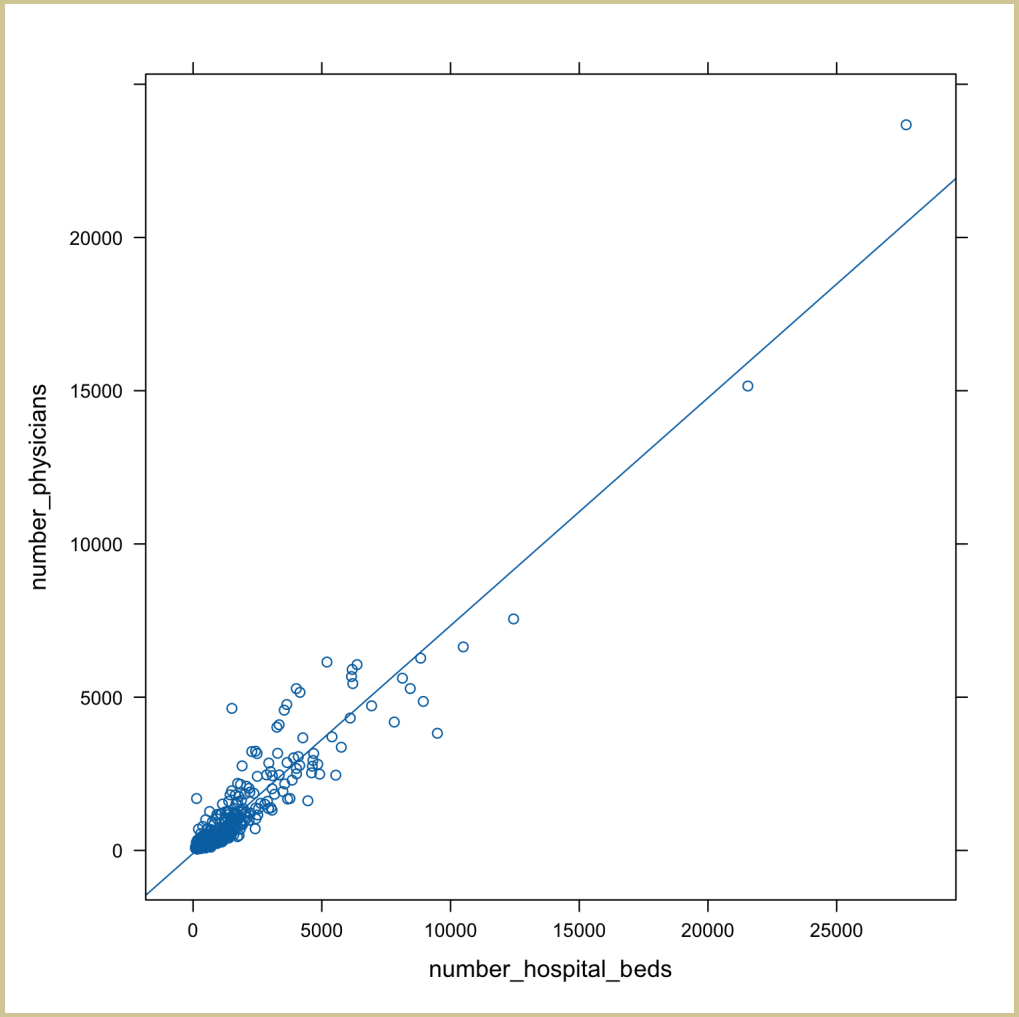
Show 20 entries Search:

	county	state	land_area	population	pop_18_to_34	pop_65	r
1	Los_Angeles	CA	4060	8863164		32.1	
2	Cook	IL	946	5105067		29.2	
3	Harris	TX	1729	2818199		31.3	
4	San_Diego	CA	4205	2498016		33.5	
5	Orange	CA	790	2410556		32.6	
6	Kings	NY	71	2300664		28.3	
7	Maricopa	AZ	9204	2122101		29.2	

Showing 1 to 20 of 440 entries

Previous 1 2 3 4 5 … 22 Next

```
mod_physician_beds = lm(number_physicians ~ number_hospital_beds, data=
xyplot(number_physicians ~ number_hospital_beds, data=cdi,  type=c("p",
```



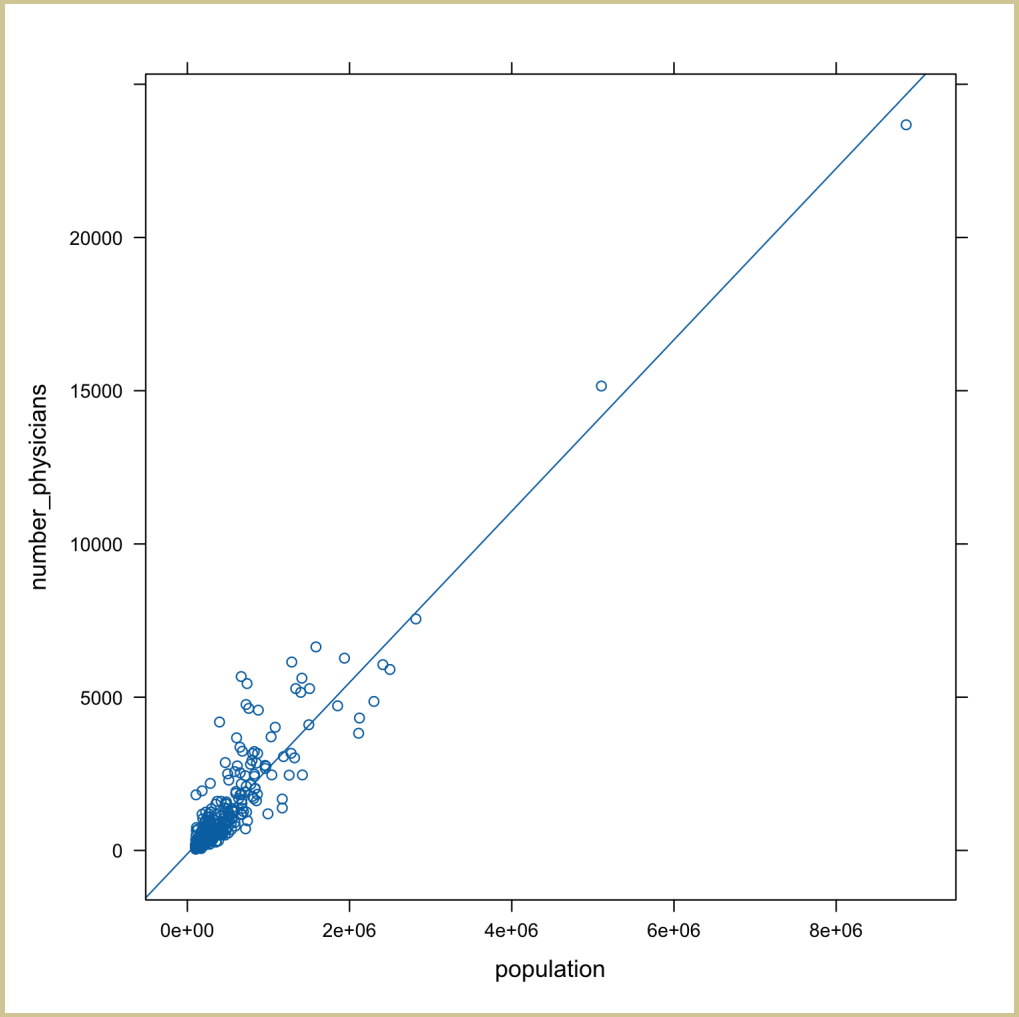
```
s = summary(mod_physician_beds)$sigma
s
```

```
## [1] 556.9487
```

```
s^2
```

```
## [1] 310191.9
```

```
mod_physician_pop = lm(number_physicians ~ population, data=cdi)
xyplot(number_physicians ~ population, data=cdi,  type=c("p", "r"))
```



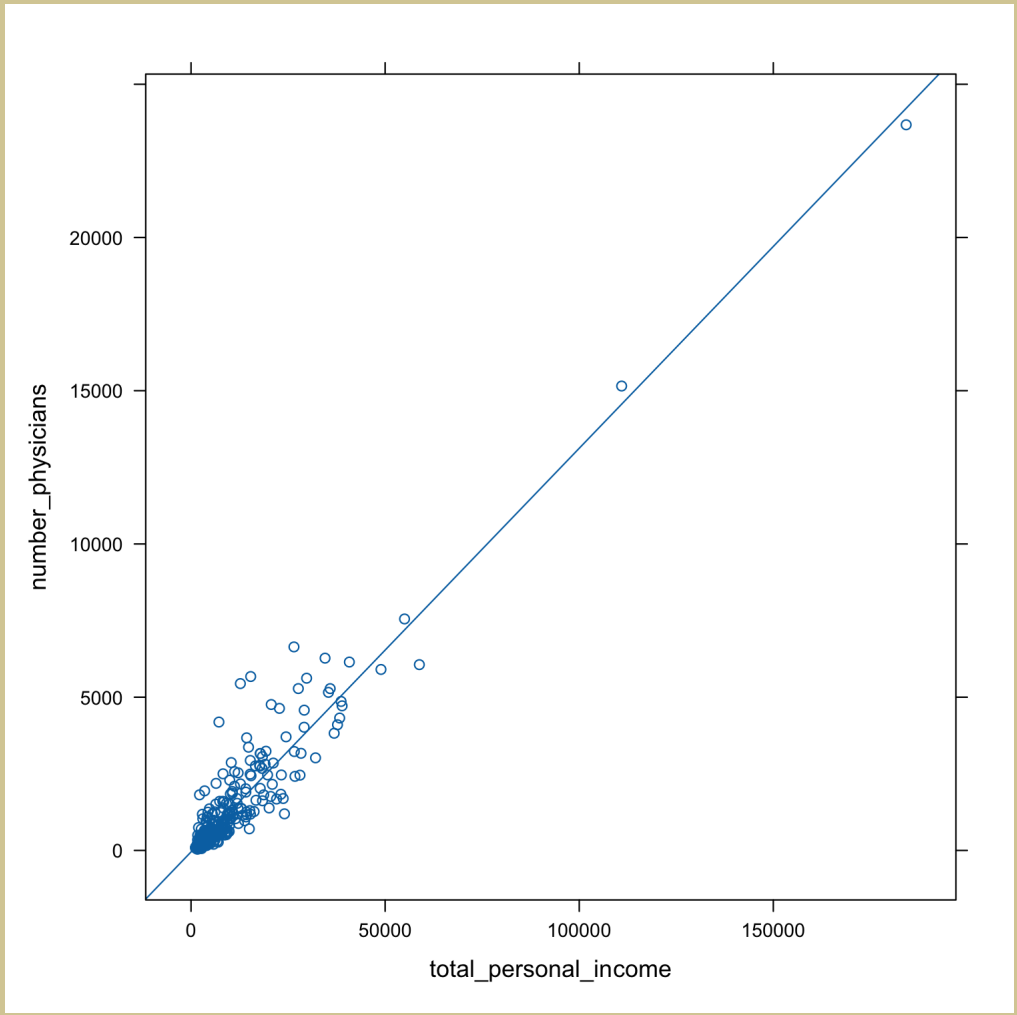
```
s = summary(mod_physician_pop)$sigma
s
```

```
## [1] 610.0848
```

```
s^2
```

```
## [1] 372203.5
```

```
mod_physician_income = lm(number_physicians ~ total_personal_income, da
xyplot(number_physicians ~ total_personal_income, data=cdi,  type=c("p"
```



```
s = summary(mod_physician_income)$sigma
s
```

```
## [1] 569.6836
```

```
s^2
```

```
## [1] 324539.4
```

```
summary(mod_physician_income)
```

```
##
## Call:
## lm(formula = number_physicians ~ total_personal_income, data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1926.6  -194.5   -66.6    44.2  3819.0
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -48.39485   31.83333   -1.52    0.129
## total_personal_income   0.13170    0.00211   62.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 569.7 on 438 degrees of freedom
## Multiple R-squared:  0.8989,    Adjusted R-squared:  0.8987
```

```
s = summary(mod_physician_income)$sigma
s
```

```
## [1] 569.6836
```

```
s^2
```

```
## [1] 324539.4
```

```
msummary(mod_physician_income)
```

```
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -48.39485   31.83333   -1.52    0.129
## total_personal_income   0.13170    0.00211   62.41   <2e-16 ***
##
## Residual standard error: 569.7 on 438 degrees of freedom
## Multiple R-squared:  0.8989,    Adjusted R-squared:  0.8987
## F-statistic:  3895 on 1 and 438 DF,  p-value: < 2.2e-16
```

1.8: Normal Error Regression Model

No matter how the error terms ε_i are distributed, the least squares method provides unbiased point estimators of β_0 and β_1

However, to set up interval estimates and make tests we need to specify the distribution of the ε_i

- We will assume that the ε_i are normally distributed:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2).$$

- Y_i is the value of the response and X_i is the predictor

- β_0 and β_1 are parameters
- ε_i is a random error term with mean $E(\varepsilon_i) = 0$ and variance $Var(\varepsilon_i) = \sigma^2$
- ε_i and ε_j are uncorrelated
- ε_i are normally distributed

We can write these final 3 points succinctly by writing $\boxed{\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)}$

Estimation of Parameters by the Method of Maximum Likelihood

We have now fully specified the probability distribution of $Y|X$ up to the parameters β_0, β_1, and σ^2

- That is, we have specified $f(Y|X; \beta_0, \beta_1, \sigma^2)$
    - so for a given set of parameters $(\beta_0, \beta_1, \sigma^2)$, we can calculate the probability of specific Y values for a given X
    - i.e., $f(Y|X; \beta_0, \beta_1, \sigma^2)$ tells us how *likely* it is that Y would come from a distribution with these specific parameters.
    - Since the data are independent, the joint probability of the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ is

$$L = f(Y_1|X_1; \beta_0, \beta_1, \sigma^2) \times \cdots \times f(Y_n|X_n; \beta_0, \beta_1, \sigma^2)$$

I have denoted this function L because it is measuring how *likely* the data set is for a specific set of parameters.

However, we want to use this function the other way around:

- We have the observed data
- what we don't have is the parameters $(\beta_0, \beta_1, \sigma^2)$

We are trying to find the parameters that gave rise to the data that we actually observed

- i.e., we want to find parameters that are most consistent with the observed data

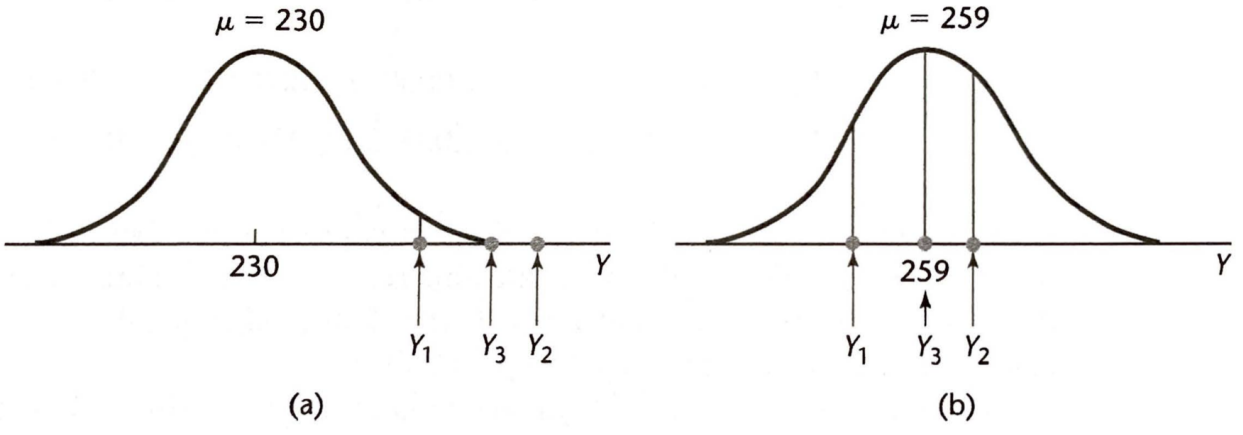
Therefore, we want to measure how *likely* the observed data would be for a given set of parameters (which we can do with $L(\beta_0, \beta_1, \sigma^2)$)

The parameter set that makes the observed data *most likely* is a good set of estimates for the unknown parameters.

Maximum Likelihood in a Single Population

- Consider a normal population with known standard deviation of $\sigma = 10$
- We observed $n = 3$ values:
  - $Y_1 = 250, Y_2 = 265, Y_3 = 259$
- We want to know what value of μ is most consistent with the sample data.

FIGURE 1.13
Densities for Sample Observations for Two Possible Values of μ: $Y_1 = 250$, $Y_2 = 265$, $Y_3 = 259$.



The densities for $Y_1 = 250$, denoted by $f_1(250; \mu)$ can be found as follows:

$$\mu = 230: \quad f_1 = \frac{1}{\sqrt{2\pi}(10)} \exp\left[-\frac{1}{2}\left(\frac{250-230}{10}\right)^2\right] = .005399$$

$$\mu = 259: \quad f_1 = \frac{1}{\sqrt{2\pi}(10)} \exp\left[-\frac{1}{2}\left(\frac{250-259}{10}\right)^2\right] = .026609$$

The densities for Y_1 can also be found as follows:

```
dnorm(250, mean=230, sd=10)
```

```
## [1] 0.005399097
```

```
dnorm(250, mean=259, sd=10)
```

```
## [1] 0.02660852
```

```
Y = c(250, 265, 259)
mu.230=dnorm(Y, mean=230, sd=10)
mu.259=dnorm(Y, mean=259, sd=10)

data.frame(Y, mu.230, mu.259) %>% round(5) %>% datatable(options=list(s
```

Show 20 entries Search:

	Y	mu.230	mu.259
1	250	0.0054	0.02661
2	265	0.00009	0.03332
3	259	0.0006	0.03989

Showing 1 to 3 of 3 entries Previous 1 Next

```
prod(mu.230)
```

```
## [1] 2.804654e-10
```

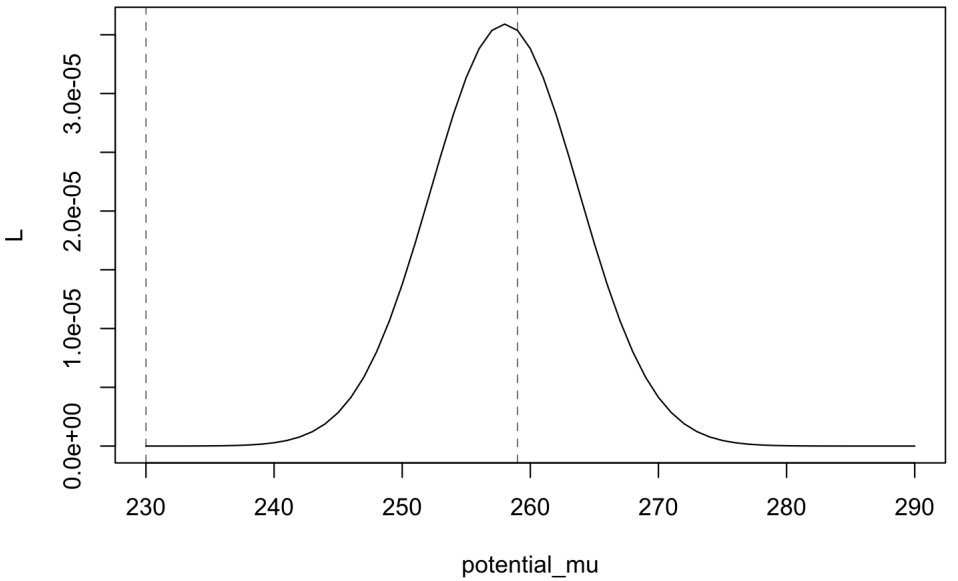
```
prod(mu.259)
```

```
## [1] 3.537268e-05
```

```r
potential_mu = 230:290
L1=dnorm(250, mean=potential_mu, sd=10)
L2=dnorm(265, mean=potential_mu, sd=10)
L3=dnorm(259, mean=potential_mu, sd=10)

L = L1 * L2 * L3

plot(potential_mu, L, type="l")
abline(v=c(230, 259), lty=2, lwd=.5)
```



Notice that the likelihood function

1. reaches its maximum at $\mu = 258$, which is the sample mean $\bar{Y} = \frac{1}{3}(250 + 265 + 259)$
2. is relatively peaked
    - so the values of μ not near the maximum likelihood estimator are much less consistent with the sample data
    - this maximum likelihood estimate is relatively *precise*

Finding the maximum likelihood analytically

Above, we found the maximum likelihood through a numerical search

- we calculated $L(\mu) = \prod_{i=1}^{n} f(Y_i; \mu)$ for a large number of possible values of μ and identified the one that resulted in the largest value

Often, however, we can find the maximum of the likelihood function analytically.

- we just need to find solutions to

$$\frac{dL(\mu)}{d\mu} = 0$$

- in practice, it is almost always better to focus on maximizing the log-likelihood

$$\frac{d\ell(\mu)}{d\mu} = \frac{d \log L(\mu)}{d\mu} = \frac{d \log\left[\prod_{i=1}^{n} f(Y_i; \mu)\right]}{d\mu} = \frac{d \sum_{i=1}^{n} \log[f(Y_i; \mu)]}{d\mu}$$

Derive the maximum of the log likelihood for μ in a normal population where σ is known:

$$\ell(\mu) = \log L(\mu) = \sum_{i=1}^{n} \log f(Y_i; \mu) = \sum_{i=1}^{n} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2} \left(\frac{Y_i - \mu}{\sigma} \right)^2 \right\} \right]$$

Maximum Likelihood in a Regression Model

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{Y_i - (\beta_0 + \beta_1 X_i)}{\sigma}\right)^2\right]$$

We can maximize this with respect to the parameters $(\beta_0, \beta_1, \sigma^2)$. (How?)

- We find that the maximum likelihood estimators are
  - $\hat{\beta}_0 = b_0$
  - $\hat{\beta}_1 = b_1$
  - $\hat{\sigma}^2 = \frac{1}{n}\sum(Y_i - \hat{Y}_i)^2$
- The maximum likelihood estimators of β_0 and β_1 are the same estimators as those provided by the method of least squares!
- The maximum likelihood estimator is biased, so ordinarily the unbiased estimator MSE is used:

$$s^2 = MSE = \frac{n}{n-2}\hat{\sigma}^2$$

Recap: Sections 1.7-1.8

After Sections 1.7-1.8, you should be able to

- Define the normal error regression model
- Define and interpret SSE and MSE
- Apply the method of maximum likelihood