Topic 11: Correlation and Linear Regression

1

Modelling Relationships

- In many applications we want to know is there a **relationship** between variables
- **Regression** is a set of statistical methods for estimating the relationship between **a response variable** and **one or more explanatory variables**
- Regression may have the aim of **explanation** (describing & quantifying relationships between variables) or **prediction** (how well can we predict a response variable from explanatory variables)
- In this section we focus on **linear relationships** between variables

2

2

Learning outcomes

After careful study of this section, you should be able to:

1. Understand correlation.
2. Use simple linear regression to model linear relationships in scientific data.
3. Define residuals and residual standard error
4. Understand how the method of least squares is used to estimate the parameters in a linear regression model.
5. Interpret the coefficients of a simple linear regression model
6. Use the regression model to make a prediction of the response variable based on the explanatory variable.
7. Confidence intervals and prediction intervals for predictions

3

3

Motivation

- Many problems in science involve exploring the relationships between two or more variables.
- Scatterplots are the best way to start observing the relationship and the ideal way to picture associations (e.g. correlation) between two *continuous* variables.
  - When the roles are clear, the explanatory or predictor variable goes on the *x*-axis, and the response variable (variable of interest) goes on the *y*-axis.
- The statistical technique known as ***Regression*** allows the researcher to *model* the dependency of a ***Response*** variable on one or more ***Explanatory*** variables.

4

Motivating Example

- Windfarms are used to generate direct current. Data are collected on 34 different days to determine the relationship between wind speed in mi/h and current in kA.

5

Data:

Name of data file: Windspeed.csv

Response Variable: current in kA

Explanatory Variable: wind speed in mi/h

6

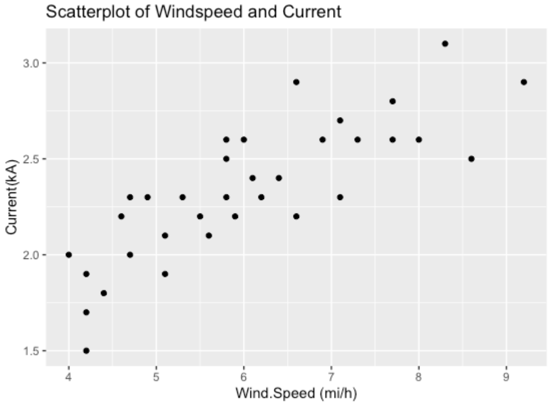
```
windspeed.df %>%
  select(Current, Wind.Speed) %>%
  summary()

##     Current         Wind.Speed
##  Min.   :1.500   Min.   :4.000
##  1st Qu.:2.125   1st Qu.:4.950
##  Median :2.300   Median :5.850
##  Mean   :2.335   Mean   :6.047
##  3rd Qu.:2.600   3rd Qu.:7.050
##  Max.   :3.100   Max.   :9.200
```

7

```
ggplot(windspeed.df, aes(y = Current, x = Wind.Speed)) +
  geom_point() +
  labs(x = "Wind.Speed (mi/h)", y = "Current(kA)",
       title = "Scatterplot of Windspeed and Current")
```

8



9

Subjective Impressions ?

- Does it look like there is a relationship between windspeed and current ?
- What is the direction of relationship ?
- How would you quantify the strength of the relationship ?

10

Sample Correlation Coefficient

The sample correlation coefficient *(r)* gives a numerical measurement of the strength of the linear relationship between the explanatory and response variables.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

11

Correlation Coefficient

$\rho = +1$ means a perfect, linear direct relationship between X and Y

$\rho = 0$ means no linear relationship between X and Y

$\rho = -1$ means a perfect, inverse linear relationship between X and Y.

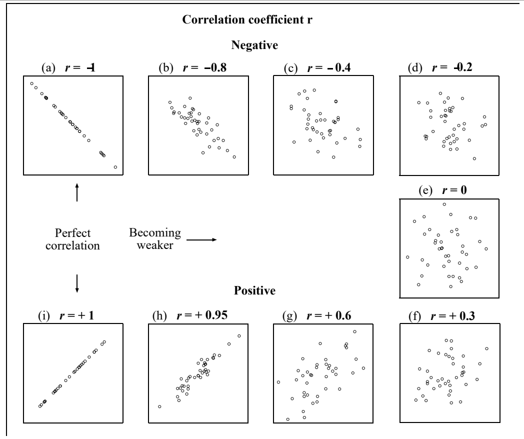
Note: ρ is the population correlation coefficient while
r is the sample correlation coefficient.

12

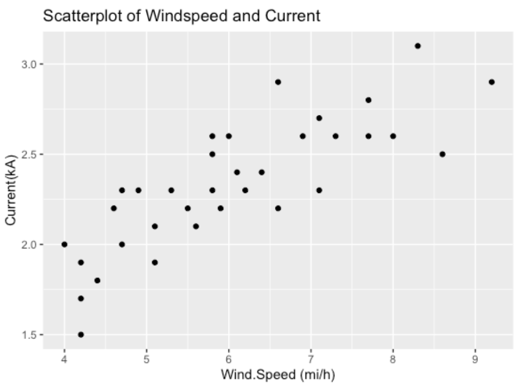
Correlation Coefficient

- Correlation treats *x* and *y* symmetrically:
  - The correlation of *x* with *y* is the same as the correlation of *y* with *x*.
- Correlation has no units.
- Correlation is not affected by changes in the center or scale of either variable.

13



14



15

Correlation coefficient

```
windspeed.df %>%
  select(Current, Wind.Speed) %>%
  cor()

##              Current Wind.Speed
## Current    1.0000000  0.8169993
## Wind.Speed 0.8169993  1.0000000
```

Or directly using **cor** function as below:

```
```{r}
cor(windspeed.df$Current, windspeed.df$Wind.Speed)
```

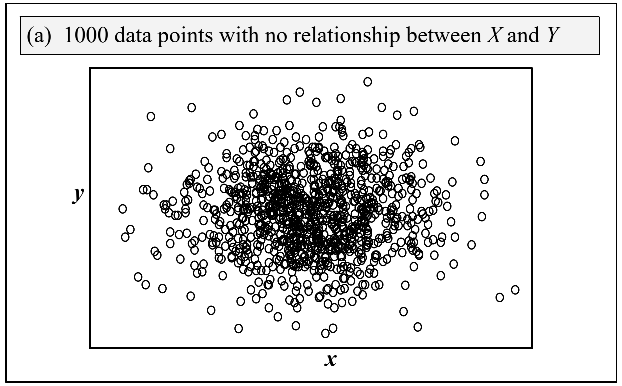
[1] 0.8169993
```

16

Correlation of zero ?

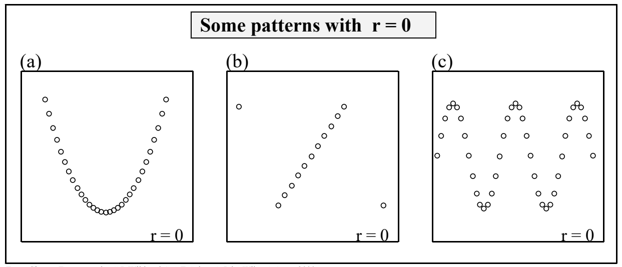
- Sketch what it looks like …

17



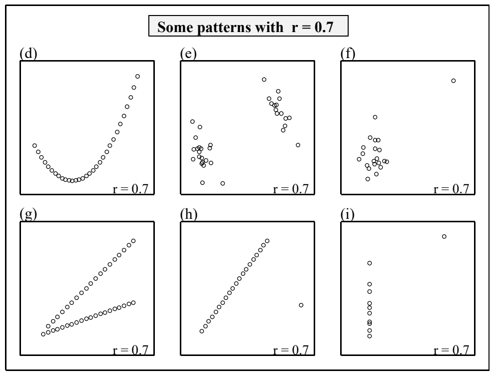
From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 1999.

18



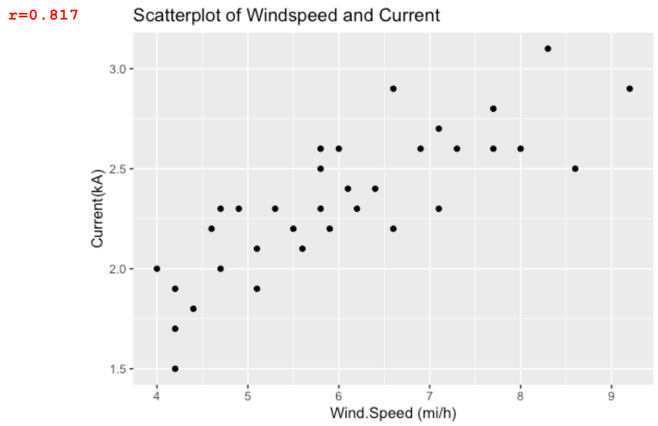
From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

19



From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

20



21

Take home message …

- Show me the data
- The correlation coefficient measures only linear association
- The correlation coefficient can be misleading when outliers are present
- **Correlation does not imply causation**

22

Correlation ≠ Causation

- Whenever we have a strong correlation, it is tempting to explain it by imagining that the predictor variable has caused the response to help.
- Scatterplots and correlation coefficients never prove causation.
- A hidden variable that stands behind a relationship and determines it by simultaneously affecting the other two variables is called a lurking or confounding variable.

23

Correlation ≠ Causation

- Don’t say “correlation” when you mean “association.
- More often than not, people say correlation when they mean association.
- The word “correlation” should be reserved for measuring the strength and direction of the linear relationship between two quantitative variables.

24

Summary so far ....

- Scatterplots are useful graphical tools for assessing *direction*, *form*, *strength*, and *unusual features* between two variables.
- Although not every relationship is linear, when the scatterplot is straight enough, the *correlation coefficient* is a useful numerical summary.
  - The sign of the correlation tells us the direction of the association.
  - The magnitude of the correlation tells us the *strength* of a linear association.
  - Correlation has no units, so shifting or scaling the data, standardizing, or swapping the variables has no effect on the numerical value.

25

Simple Linear Regression

- Simple linear regression is the name given to the statistical technique that is used to model the dependency of a response variable on a single explanatory variable
  - the word 'simple' refers to the fact that a single explanatory variable is available.
- Simple linear regression is appropriate if the average value of the response variable is a ***linear*** function of the explanatory i.e. the underlying dependency of the response on the explanatory appears linear.

26

Strategy

- Propose a model
- Check the assumptions
- Make some predictions
- Assess how useful it is
- Improve it.

27

Simple Linear Regression

OpenIntro Statistics

Fourth Edition

28

28

ModernDive

6 Basic Regression

Now that we are equipped with data visualization skills from Chapter 3, an understanding of the "tidy" data format from Chapter 4, and data wrangling skills from Chapter 5, we now proceed with data modeling. The fundamental premise of data modeling is *to make explicit the relationship* between:

- an outcome variable y, also called a dependent variable and
- an explanatory/predictor variable x, also called an independent variable or covariate.

29

Motivating Example

- Windfarms are used to generate direct current. Data are collected on 34 different days to determine the relationship between wind speed in mi/h and current in kA.

30

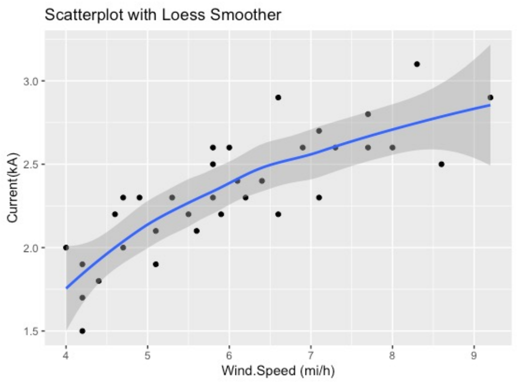
A glimpse of the first few rows of data ..

	Wind.Speed <dbl>	Current <dbl>
1	4.2	1.9
2	6.6	2.2
3	4.7	2.0
4	5.8	2.6
5	5.8	2.3
6	7.3	2.6
7	7.1	2.7
8	6.4	2.4
9	4.6	2.2
10	4.2	1.5

31

```
ggplot(windspeed.df, aes(y = Current, x = Wind.Speed)) +
  geom_point() +
  geom_smooth() +
  labs(x = "Wind.Speed (mi/h)", y = "Current(kA)",
       title = "Scatterplot with Loess Smoother")
```

32



33

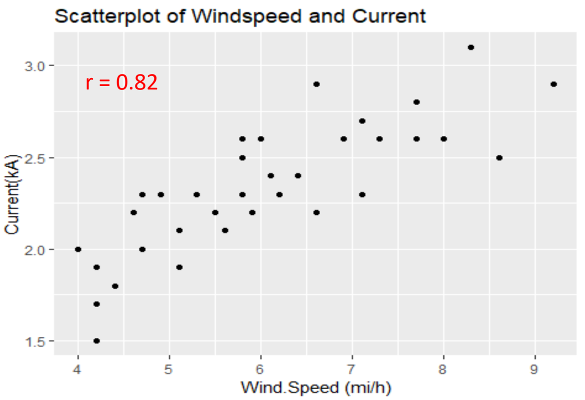
Simple Linear Regression

- The simple linear regression model is of the form

Response Variable = Intercept + Slope*Explanatory Variable
+ random variability

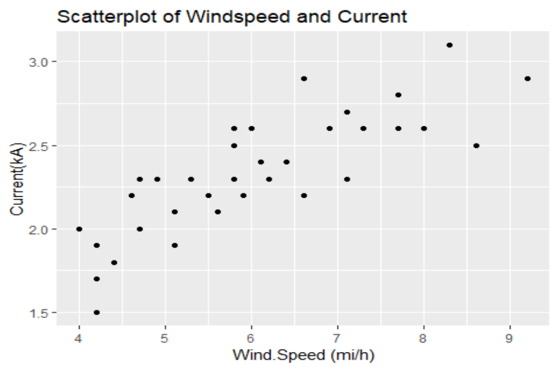
where the intercept and slope must be estimated from a relevant sample of data from the population of interest.

34



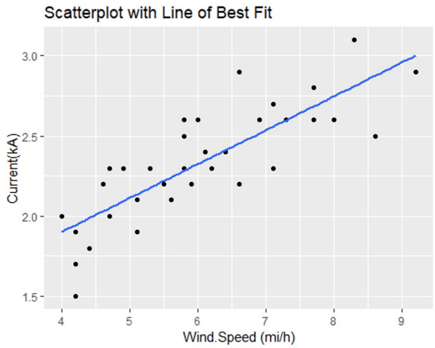
35

Line of best fit ?



36

```r
ggplot(windspeed.df, aes(y=Current, x=Wind.Speed)) +
  geom_point() +
geom_smooth(method = "lm", se= FALSE) +
  labs(x = "Wind.Speed (mi/h)", y = "Current(kA)",
       title = "Scatterplot with Line of Best Fit")
```



37

Interpreting the Slope and Intercept

<u>Regression Equation</u>
Mean Current = 1.057 + 0.2113 Wind Speed

- b_1 is the slope, which tells us how rapidly $\hat{y}$ changes with respect to x e.g. what is the change in the mean current per unit increase in wind speed.
- b_0 is the *y*-intercept, which tells where the line crosses (intercepts) the y-axis when x is zero e.g. what is the mean current when wind speed is zero.

38

Predict the Current when Wind Speed = 7.1

<u>Regression Equation</u>
Mean Current = 1.057 + 0.2113 Wind Speed

39

Predict the Current when Wind Speed = 7.1

<u>Regression Equation</u>
Mean Current = 1.057 + 0.2113 (7.1) = 2.56

The predicted value is often referred to as $\hat{y}$ (i.e. 'y hat').

From looking at the data the 7th observation was for a wind speed of 7.1 where the *actual* Current (i.e. y) was equal to 2.7.

40

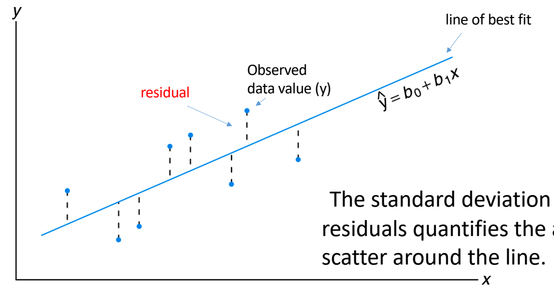
Residuals …. difference between actual and predicted values

	Wind.Speed <dbl>	Current <dbl>
1	4.2	1.9
2	6.6	2.2
3	4.7	2.0
4	5.8	2.6
5	5.8	2.3
6	7.3	2.6
7	7.1	2.7
8	6.4	2.4
9	4.6	2.2
10	4.2	1.5

Actual current = 2.7
Predicted current ($\hat{y}$)= 2.56
The difference (Actual – Predicted) = 0.14

41

The line of best fit is the line for which the sum of the squared residuals is smallest, the least squares line.



The standard deviation s_e of the residuals quantifies the amount of scatter around the line.

42

Actual current

Predicted current

Actual - Predicted

```
get_regression_points(windspeed.model)

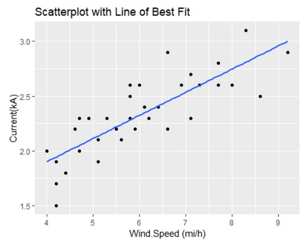
## # A tibble: 34 x 5
##       ID Current Wind.Speed Current_hat residual
##    <int>   <dbl>      <dbl>       <dbl>    <dbl>
## 1      1     1.9        4.2        1.94   -0.045
## 2      2     2.2        6.6        2.45   -0.252
## 3      3     2          4.7        2.05   -0.051
## 4      4     2.6        5.8        2.28    0.317
## 5      5     2.3        5.8        2.28    0.017
## 6      6     2.6        7.3        2.6     0
## 7      7     2.7        7.1        2.56    0.142
## 8      8     2.4        6.4        2.41   -0.01
## 9      9     2.2        4.6        2.03    0.171
## 10    10     1.5        4.2        1.94   -0.445
## # ... with 24 more rows
```

s_e
standard deviation of the residuals

43

The Residual Standard Deviation (s_e)

- The standard deviation of the residuals, s_e, measures how much the points spread around the regression line.
- Also known as the residual standard error.
- You can interpret s_e in the context of a data set. It is the typical error in the predictions made by the regression line.



44

Line of 'best fit'.

- The line of best fit is the line for which the sum of the squared residuals is smallest, the least squares line.
- Some residuals are positive, others are negative, and, on average, they cancel each other out.
- You can't assess how well the line fits by adding up all the residuals.

45

- Simple Linear Regression Model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{for } i=1, \ldots, n \text{ assuming } \varepsilon_i \sim N(0, \sigma_e)$$

- Features of this model:
- β_0 (intercept) and β_1 (slope) are the population parameters of the model and must be estimated from the data as b_0 (sample intercept) and b_1 (sample slope).
- The process of estimating β_0 and β_1 is called fitting the model to the data.
- $\beta_0 + \beta_1 x_i$ is the population mean response (mean of Y) given $X=x_i$.
- ε_i is the error term in the regression model. Actually it refers to the difference between the fitted line and y_i.
- σ_e (error) is the stochastic part of the model (unexplained variability). Or in other words, it is the standard deviation corresponding to the error term.
- Once estimated predicted values for y (labelled as $\hat{y}$) can be made as follows:

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

- $\hat{y}$ is used to emphasize that the points that satisfy this equation are just our *predicted* values, not the actual data values.

46

Estimating the Slope (least squares)

- In the simple linear regression model the slope (b_1) is built from the correlation coefficient r and the standard deviations of y and x:

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

- The slope is always in units of *y* per unit of *x*.

47

Estimating the Intercept (least squares)

- In the simple linear regression model the intercept (b_0) the intercept is built from the means and the slope:

$$b_0 = \bar{y} - b_1 \bar{x}$$

- The intercept is always in units of *y*.
- We almost always use technology to find the equation of the regression line.

48

Summary Statistics

```
windspeed.df %>%
              summarize(Mean.Current=mean(Current), SD.Current= sd(Current),
                 Mean.Windspeed=mean(Wind.Speed), S.Windspeed= sd(Wind.Speed))

##   Mean.Current SD.Current Mean.Windspeed S.Windspeed
## 1     2.335294  0.3583484       6.047059    1.385255
```

````
```{r}
cor(windspeed.df$Current, windspeed.df$Wind.Speed)
```
````

```
[1] 0.8169993
```

49

Slope and Intercept

$b_1 = r\frac{s_y}{s_x} =$,

$b_0 = y - b_1 x =$

Regression Equation
Mean Current = + Wind Speed

50

Slope and Intercept

$b_1 = r\frac{s_y}{s_x} = 0.817\frac{0.3583}{1.385} = 0.2113,$

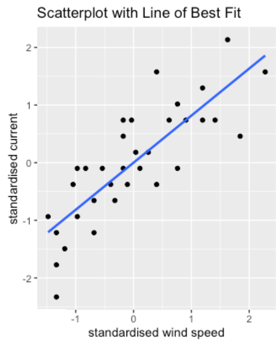
$b_0 = y - b_1 x = 2.3353 - 0.2113(6.047) = 1.057$

Regression Equation
Mean Current = 1.057 + 0.2113 Wind Speed

51

Windspeed and current standardized (subtract mean and divide by sd)
So standardized values have mean 0 and sd 1

r = 0.817




```
> lm(currentstd ~ windspeedstd, data = windspeed.df)

Call:
lm(formula = currentstd ~ windspeedstd, data = windspeed.df)

Coefficients:
 (Intercept)  windspeedstd
  -5.052e-16     8.170e-01
```

52

Summary so far …

- **Correlation** is a useful metric for measuring the degree of **linear relationship** between two continuous variables
- **Regression** is a useful tool for **modelling** the relationship between two continuous variables: a response (y) and an explanatory/predictor (x)
- The line of best fit is the line where the sum of the squared residuals (difference between observed and fitted values) is a minimum
- To use this line to make **inference** (and predictions) there are several **assumptions** that must be satisfied

53

Fitting a Simple Linear Regression in R

```
```{r}
windspeed.model <- lm(Current ~ Wind.Speed, windspeed.df)
windspeed.model
```
```

```
Call:
lm(formula = Current ~ Wind.Speed, data = windspeed.df)

Coefficients:
(Intercept)   Wind.Speed
     1.0573       0.2113
```

54

Fitting a Simple Linear Regression in R

```
```{r}
windspeed.model <- lm(Current ~ Wind.Speed, windspeed.df)
windspeed.model
```
```

```
Call:
lm(formula = Current ~ Wind.Speed, data = windspeed.df)

Coefficients:
(Intercept)   Wind.Speed
     1.0573       0.2113
```

Intercept

Slope

55

Interpreting the Slope and Intercept

```
Regression Equation
Mean Current = 1.057 + 0.2113 Wind Speed
```

- b_1 is the slope, which tells us how rapidly $\hat{y}$ changes with respect to x e.g. what is the change in the (mean) current per unit increase in wind speed.
- b_0 is the y-intercept, which tells where the line crosses (intercepts) the y-axis when x is zero e.g. what is the (mean) current when wind speed is zero.

56

Inference for predictions

- We have seen how to make **point estimates** of the predicted response
- Just as in inference for the true mean, an interval estimate is more useful for inference
- We look at two types of **interval estimates for the mean (or predicted) response given some value of the explanatory variable**
- 1. Confidence interval
- 2. Prediction interval

57

Confidence Interval for the mean response

- A range of values that is likely to contain the **true mean value of the response variable given a specific values of the the explanatory variable**.
- This range **doesn't tell** you about the spread of the **individual data points** around the true mean.

58

Prediction Interval for response in new observations

- A range of values that is likely to contains the value of the response variable for a **single new observation** given a specific value of the explanatory variable.
- The prediction interval is for **individual observations rather than the mean**.

59

For prediction in R: the predict() function

- predict(object, newdata, se.fit = FALSE, interval = c("none", "confidence", "prediction"), level = 0.95)
- object a fitted lm() model object.
- newdata An optional data frame in which to look for variables with which to predict.
- se.fit A switch indicating if standard errors for predictions are required. The default is se.fit = FALSE.
- interval Type of interval to be calculated. The default is interval = "none".
- level the confidence level for generating interval estimates. The default is level = 0.95.

60

R code for confidence interval and prediction interval for a single point

```
> fit<-lm(Current ~ Wind.Speed, data = windspeed.df)
> new.d <- data.frame(Wind.Speed = 7)
> predict(fit, newdata = new.d, interval = "confidence", level = 0.95)
       fit     lwr      upr
1 2.536696 2.44729 2.626102
> predict(fit, newdata = new.d, interval = "prediction", level = 0.95)
       fit      lwr      upr
1 2.536696 2.100013 2.973379
> 
```

61

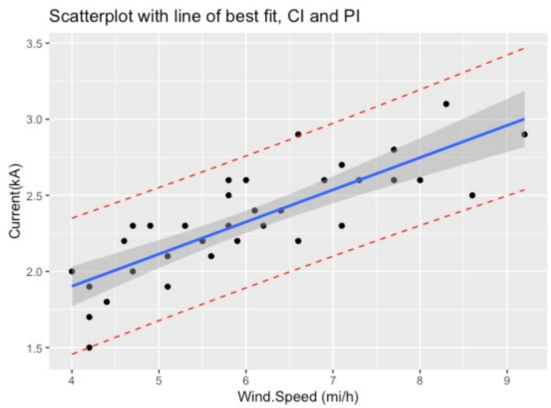
R code for pointwise CI and PI

```
pred.int <-  predict(fit, newdata = windspeed.df, interval = "prediction")

windspeed.df2 <- cbind(windspeed.df, pred.int)

windspeed.df2 %>%
    ggplot(aes(x = Wind.Speed, y = Current)) +
    geom_point() +
    stat_smooth(method = lm) +
    geom_line(aes(y = lwr), color = "red", linetype = "dashed") +
    geom_line(aes(y = upr), color = "red", linetype = "dashed") +
    labs(x = "Wind.Speed (mi/h)", y = "Current(kA)",
         title = "Scatterplot with line of best fit and Confidence and Prediction Interv
ıls")
```

62



63

What Can Go Wrong?

- Don't fit a straight line to a nonlinear relationship.
- Beware extraordinary points (**y-values that stand off from the linear pattern or extreme *x*-values**).
- Don't extrapolate beyond the data—the linear model may no longer hold outside of **the range of the data**.
- Don't infer that *x* causes *y* just because there is a good linear model for their relationship—association is *not* causation.
- An empirical model is valid only for the data to which it is fit. It may or may not be useful in predicting outcomes for subsequent observations.

64

Exam Tips

Make sure you can find the following values from a computer's regression output:

1. The explanatory and response variables
2. The corresponding regression equation by finding intercept and slope.
3. Use the equation to predict for a new value of explanatory variable.

65