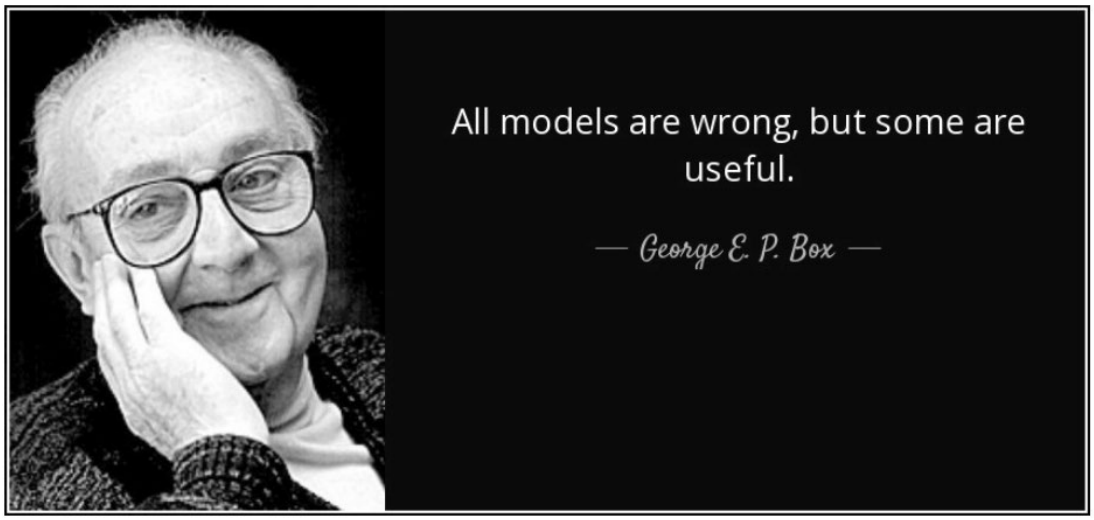
Notes and Exercises Chapters 7 – Linear Regression
Math 217
Saidi



Chapter 7

Linear regression assumes that the relationship between two variables, x and y, can be modeled by a straight line with some error:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Where β_0 and β_1 represent the two model parameters. These β_0 and β_1 linear model parameters are estimated using data.

- x - the explanatory or predictor variable
- y - the response variable
- ε – the error in the model (We often drop the ε term when writing down the model since our main focus is often on the prediction of the average outcome)

Example 1 Using linear regression to predict possum total body length

Brushtail possums are marsupials that live in Australia. Researchers captured 104 of these animals and took body measurements before releasing the animals back into the wild. We consider two of these measurements: the total length of each possum, from head to tail, and the length of each possum's head. The data are found in the openintro datasets.

The following comes from my Chapter 7 QMD file.

Chapter 7 Intro to Linear Regression

Work with the possum data found in the OpenIntro datasets

Hanging baby possum: https://www.scenichudson.org/viewfinder/the-opossum-a-surprising-n-y-fan-fave/

Load tidyverse, set the working directory, and read in the possum data

```
library(tidyverse)
library(tidymodels)
setwd("C:/Users/rsaidi/Dropbox/Rachel/MontColl/Datasets/Datasets")
possum <- read_csv("possum_openintro.csv")
```

Explore the possum dataset variables

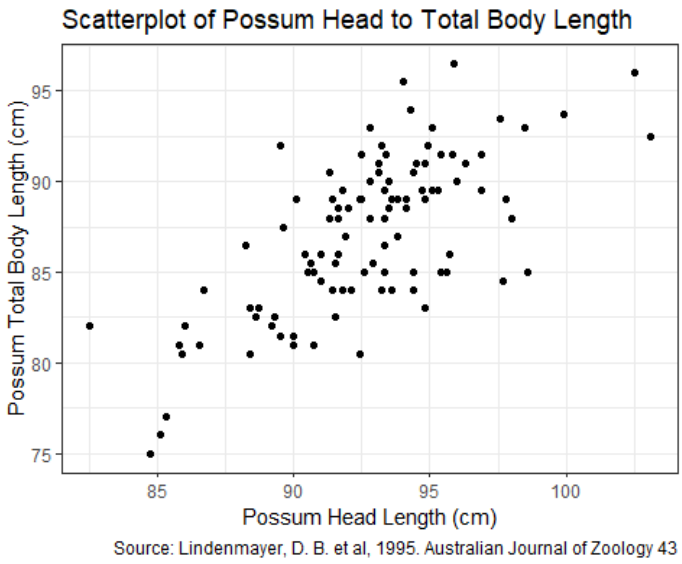
note that there are 104 observations with 8 variables. All variables are quantitative except “pop”

```
head(possum)
```

```
# A tibble: 6 × 8
   site pop   sex     age head_l skull_w total_l tail_l
  <dbl> <chr> <chr> <dbl>  <dbl>   <dbl>   <dbl>  <dbl>
1     1 Vic   m         8   94.1    60.4    89     36
2     1 Vic   f         6   92.5    57.6    91.5   36.5
3     1 Vic   f         6   94      60      95.5   39
4     1 Vic   f         6   93.2    57.1    92     38
5     1 Vic   f         2   91.5    56.3    85.5   36
6     1 Vic   f         1   93.1    54.8    90.5   35.5
```

Create a scatterplot of head length to total body length

```
ggplot(possum, aes(x=head_l, y=total_l))+
  geom_point() +
  theme_bw()+
  labs(x="Possum Head Length (cm)",
       y="Possum Total Body Length (cm)",
       title = "Scatterplot of Possum Head to Total Body Length",
       caption = "Source: Lindenmayer, D. B. et al, 1995. Australian Journal of Zoology 43")
```



We can see a positive association: as head length increases, total body length tends to increase as well.

Create a linear model to show this association

lm stands for linear model. The linear model form is: y ~ x

```
fit1 <- lm(data = possum, total_l ~ head_l)
summary(fit1)
Call:
lm(formula = total_l ~ head_l, data = possum)

Residuals:
    Min      1Q  Median      3Q     Max
-7.0881 -2.2935  0.2888  2.0801  7.4983

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.88823    8.00016   1.236    0.219
head_l       0.83367    0.08633   9.657 4.68e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.131 on 102 degrees of freedom
Multiple R-squared:  0.4776,    Adjusted R-squared:  0.4725
F-statistic: 93.26 on 1 and 102 DF,  p-value: 4.681e-16
```

We can see that this linear equation is:

$$\hat{y} = mx + b, \quad and\ in\ this\ model, \quad \widehat{total\_l} = 9.888 + 0.8336(head_l)$$

- This means that for each additional cm increase in head length, the total body length increases by 0.8337 cm.
- The y-intercept is 9.8882 cm, meaning that when the head length is 0 cm, the total body length is 9.8882 cm.

- The p-value for head_l to predict total_l is very small, and therefore head_l is an important predictor of total_l.
- The p-value for the overall model is very small, so the model is meaningful.

Finally, the adjusted R^2 is .4725. We will discuss the meaning of this value below.

Adjusted R^2

$0 \leq r^2 \leq 1$

The adjusted R^2 value may be interpreted as follows:

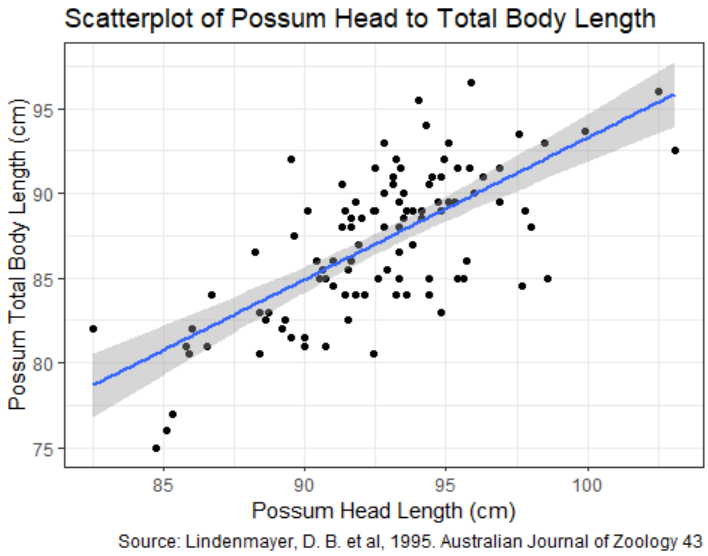
____% of the variation in the observations may be explained by this model.

For this model, 47% of the variation in the observations may be explained by this model.

Show the linear regression line on the scatterplot

```
ggplot(possum, aes(x=head_l, y=total_l))+
  geom_point() +
  geom_smooth(method = "lm") +
  theme_bw()+
  labs(x="Possum Head Length (cm)",
       y="Possum Total Body Length (cm)",
       title = "Scatterplot of Possum Head to Total Body Length",
       caption = "Source: Lindenmayer, D. B. et al, 1995. Australian Journal of Zoology 4
3")
```

```
`geom_smooth()` using formula = 'y ~ x'
```

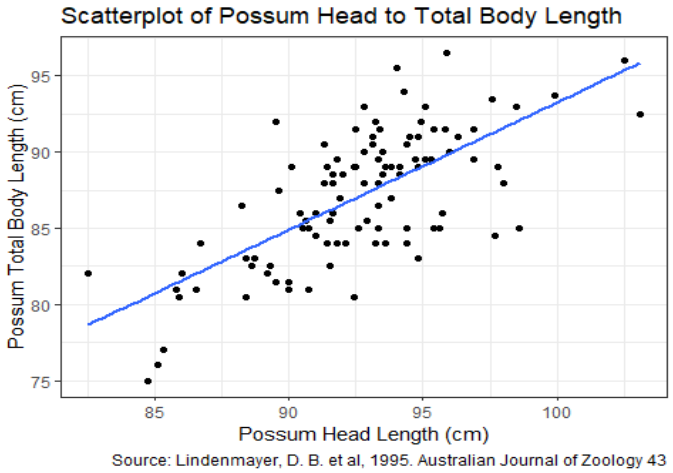


The Gray Band

The gray area around the linear regression line is the standard error (se) band. We can remove it with: "se = FALSE"

```
ggplot(possum, aes(x=head_l, y=total_l))+
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()+
  labs(x="Possum Head Length (cm)",
       y="Possum Total Body Length (cm)",
       title = "Scatterplot of Possum Head to Total Body Length",
       caption = "Source: Lindenmayer, D. B. et al, 1995. Australian Journal of Zoology 4
3")
```

```
`geom_smooth()` using formula = 'y ~ x'
```

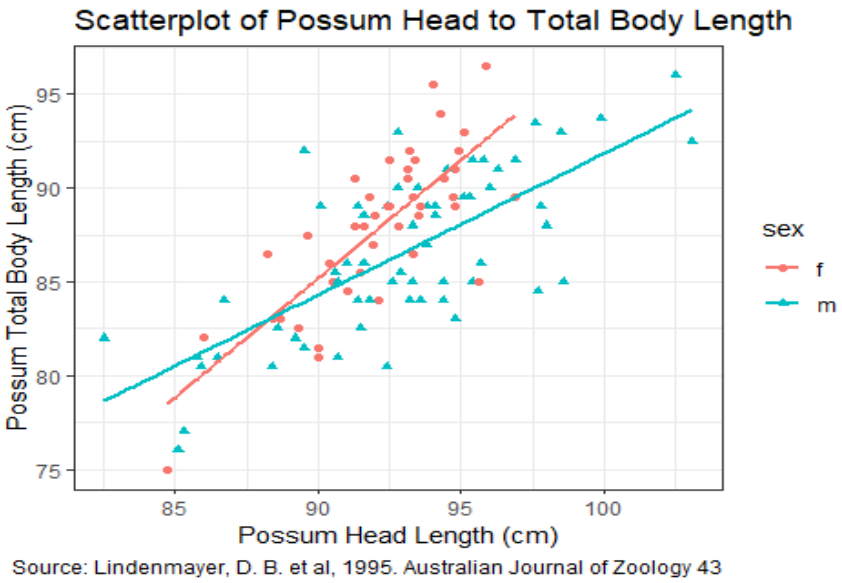


What if the relationships are different between male and female possums?

Add color and shape for the sex to distinguish males and females (sex is a categorical variable)

```
ggplot(possum, aes(x=head_l, y=total_l, color = sex))+
  geom_point(aes(shape = sex)) +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()+
  labs(x="Possum Head Length (cm)",
       y="Possum Total Body Length (cm)",
       title = "Scatterplot of Possum Head to Total Body Length",
       caption = "Source: Lindenmayer, D. B. et al, 1995. Australian Journal of Zoology 4
3")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



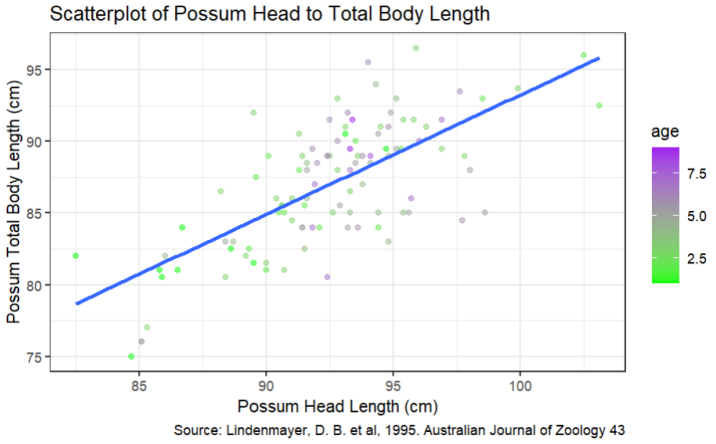
We can explore the values for the equation of this more complex plot, but just looking at the lines, the female line has a much steeper slope, indicating that as female head length increase, there is a much greater female total body length increase than for males.

Explore how age is associated with the two variables

Add color by age (which is a quantitative variable). Add alpha (the level of transparency). Alpha values go between zero and 1. Values closer to zero are MORE transparent.

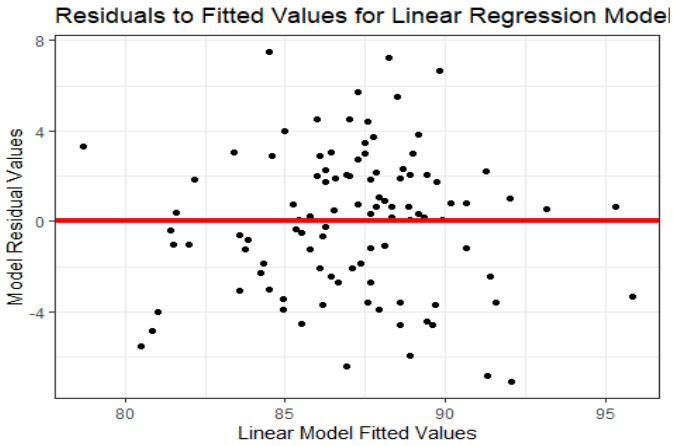
```
ggplot(possum, aes(x=head_l, y=total_l))+
  geom_point(aes(color = age), alpha = 0.5) +
  scale_color_gradient(low = "green",high = "purple",)+
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()+
  labs(x="Possum Head Length (cm)",
       y="Possum Total Body Length (cm)",
       title = "Scatterplot of Possum Head to Total Body Length",
       caption = "Source: Lindenmayer, D. B. et al, 1995. Australian Journal of Zoology 4
3")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Finally, plot the residuals

```
ggplot(data = fit1,  aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red", linewidth = 1.5) +
  theme_bw() +
  labs(x="Linear Model Fitted Values", y="Model Residual Values",
       title = "Residuals to Fitted Values for Linear Regression Model")
```



What does this residuals plot show?

The residuals plot should show points scattered in no particular pattern about the horizontal line at y=0.

Visualizing parallel slopes models

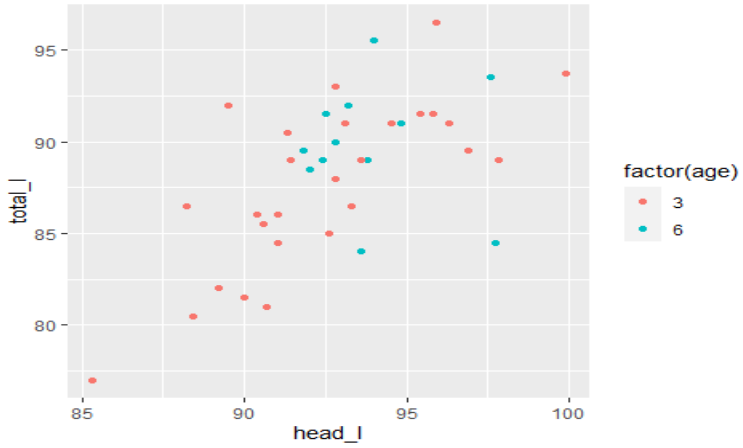
Three variables, one plot

In this scatterplot, we use color to differentiate the possums age 3 from age 6.

In this manner, we have depicted three variables—two numeric and one categorical—on the same scatterplot. Thus, this plot will enable us to visualize our parallel slopes model in the data space.

```
newdata <- possum |>
  filter(age %in% c(3, 6))

plotage1 <- ggplot(data = newdata, aes(
  x = head_l, y = total_l, color = factor(age))) +
  geom_point()
plotage1
```



Create the linear model

```
mod <- lm(total_l ~ head_l + factor(age), data = newdata)
summary(mod)


Call:
lm(formula = total_l ~ head_l + factor(age), data = newdata)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7127 -1.9868  0.2905  1.9004  6.9287

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.5115    17.4033   0.374     0.71
head_l         0.8778     0.1880   4.669 4.11e-05 ***
factor(age)6   0.9437     1.2089   0.781     0.44
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.408 on 36 degrees of freedom
Multiple R-squared:  0.4101,    Adjusted R-squared:  0.3773
F-statistic: 12.51 on 2 and 36 DF,  p-value: 7.492e-05
```

Our model is:

total_length =β_0 + β_1(head_length) + β_2(age_6)

total_length =6.5115 + 0.8778(head_length) + 0.9437(age_6)

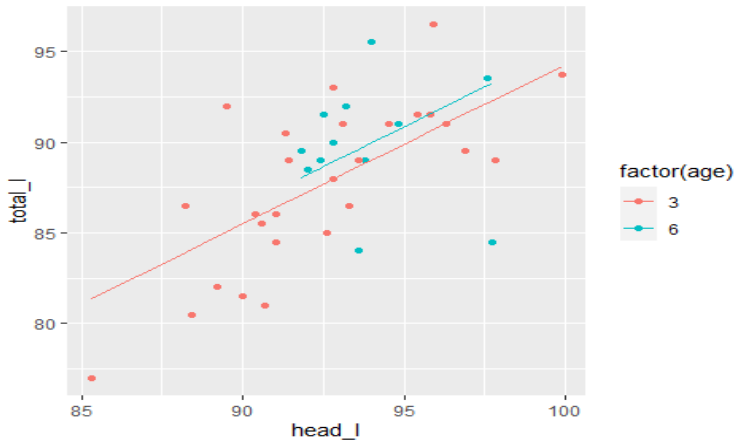
Now use the function, augment, to provide a detailed statistical summary of the model values including the fitted y-values and the residuals.

```
augment(mod)
```

```
# A tibble: 39 × 9
   total_l head_l `factor(age)` .fitted  .resid   .hat .sigma .cooksd .std.resid
     <dbl>  <dbl> <fct>           <dbl>   <dbl>  <dbl>  <dbl>   <dbl>      <dbl>
 1    91.5   92.5 6                88.6  2.85   0.0889   3.42 2.50e-2    0.877
 2    95.5   94   6                90.0  5.54   0.0834   3.31 8.73e-2    1.70
 3    92     93.2 6                89.3  2.74   0.0846   3.42 2.17e-2    0.840
 4    91     94.8 6                90.7  0.333  0.0861   3.46 3.28e-4    0.102
 5    89.5   91.8 6                88.0  1.47   0.0961   3.45 7.26e-3    0.453
 6    91.5   95.4 3                90.3  1.25   0.0625   3.45 3.19e-3    0.379
 7    96.5   95.9 3                90.7  5.81   0.0721   3.30 8.11e-2    1.77
 8    91     96.3 3                91.0 -0.0401 0.0808   3.46 4.42e-6   -0.0123
 9    91.5   95.8 3                90.6  0.899  0.0700   3.45 1.88e-3    0.274
10    88     92.8 3                88.0  0.0321 0.0373   3.46 1.19e-6    0.00959
# ℹ 29 more rows
```

Now plot the parallel slopes for ages 3 and 6

```
plotage2 <- ggplot(data = newdata, aes(
  x = head_l, y = total_l, color = factor(age))) +
  geom_point() +
  geom_line(data = augment(mod), aes(y = .fitted, color = `factor(age)`))
plotage2
```

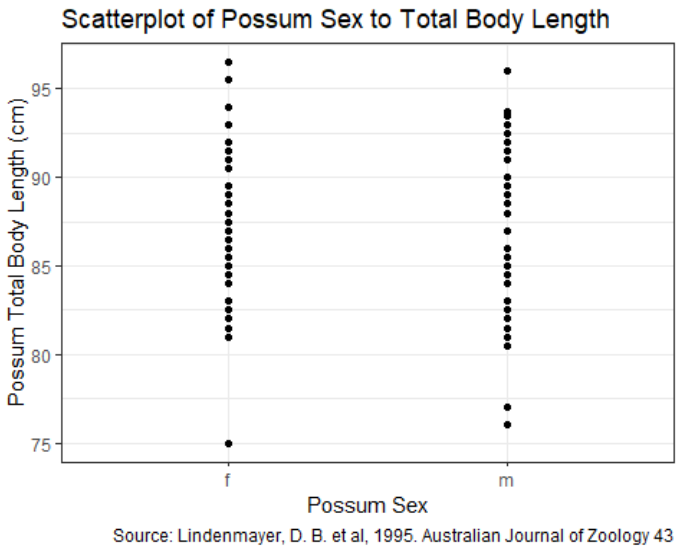


7.2.6 Categorical predictors with two levels

Using categorical predictors with only 2 factors

What if we only wanted to predict total body length of the possums based on sex? We can use what is know as a "dummy variable" - a categorical variable - as a predictor of total body length.

```
ggplot(possum, aes(x=sex, y=total_l))+
  geom_point()+
   theme_bw()+
  labs(x="Possum Sex",
       y="Possum Total Body Length (cm)",
       title = "Scatterplot of Possum Sex to Total Body Length",
       caption = "Source: Lindenmayer, D. B. et al, 1995. Australian Journal of Zoology 4
3")
```



You might be surprised, but we can create a linear model of this relationship

Create the linear model using sex and total body length

```
fit2 <- lm(data = possum, total_l ~ sex)
summary(fit2)
Call:
lm(formula = total_l ~ sex, data = possum)

Residuals:
    Min      1Q  Median      3Q     Max
-12.907  -2.511   0.093   2.989   9.489

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  87.9070     0.6520 134.819   <2e-16 ***
sexm         -1.3955     0.8514  -1.639    0.104
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.276 on 102 degrees of freedom
Multiple R-squared:  0.02566,   Adjusted R-squared:  0.01611
F-statistic: 2.687 on 1 and 102 DF,  p-value: 0.1043
```

What happened in this model?

Notice that the model ONLY SHOWS sex_m, meaning the male slope. This means that because female is alphabetically before males, R coded females as zeros and males as ones. sex_f (females): the slope for female relationship to body length is embedded in the intercept. This is called a **reference level.**

This also means that, because the male slope is negative, males total body length decreases by 1.4 cm as compared to female body length.

You may also notice that the Adj R^2 value is 0.016, which means that only 1.6% of the variation in the observations may be explained by this model – sex is not a good predictor of total body length.

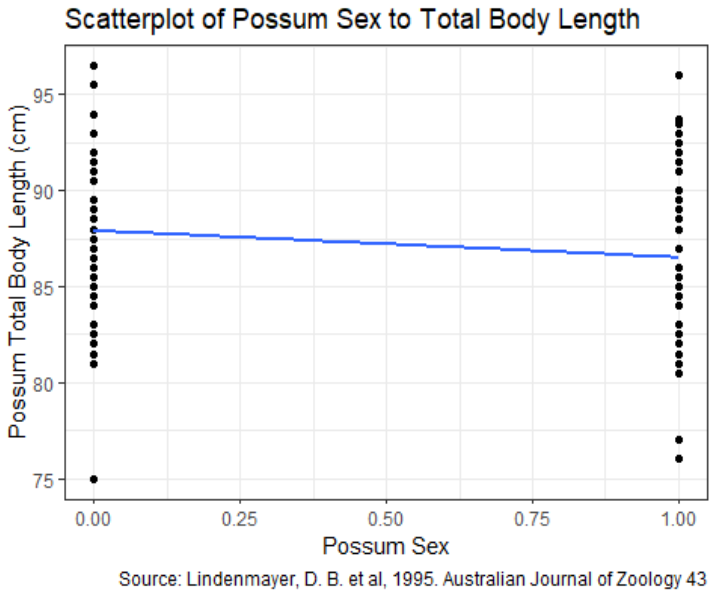
Visualize this with a plot and linear regression line

Recode sex such that male = 1 and female = 0

```
possum$sex2<-ifelse(possum$sex=="m", 1, 0)

ggplot(possum, aes(x=sex2, y=total_l))+
  geom_point()+
  geom_smooth(method = "lm", se = FALSE)+
   theme_bw()+
  labs(x="Possum Sex",
       y="Possum Total Body Length (cm)",
       title = "Scatterplot of Possum Sex to Total Body Length",
       caption = "Source: Lindenmayer, D. B. et al, 1995. Australian Journal of Zoology 4
3")

`geom_smooth()` using formula = 'y ~ x'
```



Ignoring head length and only using sex as a predictor, males overall have shorter body lengths than females.

What is very important to notice, is that one single variable cannot possibly completely explain a possum's total body length. Therefore, in Chapter 8, we will explore multiple factors causing a response variable.

7.1.4 Describing Linear Relationships with the Correlation Coefficient

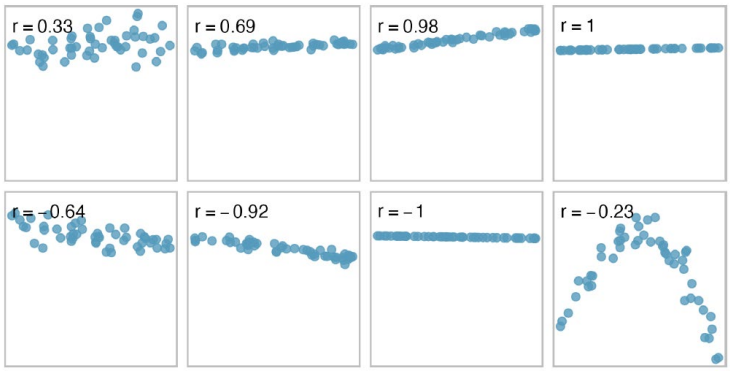
r indicates two concepts:

1. Strength $|r| \approx 1$ indicates strong correlation and $|r| \approx 0$ indicates weak to no correlation.
2. Direction of association – positive when slopes are positive and negative when slopes are negative.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

The figure below shows eight plots and their corresponding correlations. Only when the relationship is perfectly linear is the correlation either -1 or 1. If the relationship is strong and positive, the correlation will be near +1. If it is strong and

negative, it will be near -1. If there is no apparent linear relationship between the variables, then the correlation will be near zero.



7.1.3 Residuals

Residual: is the **prediction error**, is $e_i = y_i - \hat{y}_i$ (observed - predicted)

So the **residual sum of squares** is:

$$SS(residual) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} {e_i}^2$$

There is calculus involved in ***minimizing*** the RSS (using optimization techniques).

7.2.4 Extrapolation is treacherous

When those blizzards hit the East Coast this winter, it proved to my satisfaction that global warming was a fraud. That snow was freezing cold. But in an alarming trend, temperatures this spring have risen. Consider this: On February 6 it was 10 degrees. Today it hit almost 80. At this rate, by August it will be 220 degrees. So clearly folks the climate debate rages on. (Stephen Colbert April 6th, 2010)

Linear models can be used to approximate the relationship between two variables. However, like any model, they have real limitations. Linear regression is simply a modeling framework. The truth is almost always much more complex than a simple line. For example, we do not know how the data outside of our limited window will behave.

7.3 Outliers in linear regression

In this section, we identify criteria for determining which outliers are important and influential. Outliers in regression are observations that fall far from the cloud of points. These points are especially important because they can have a strong influence on the least squares line.

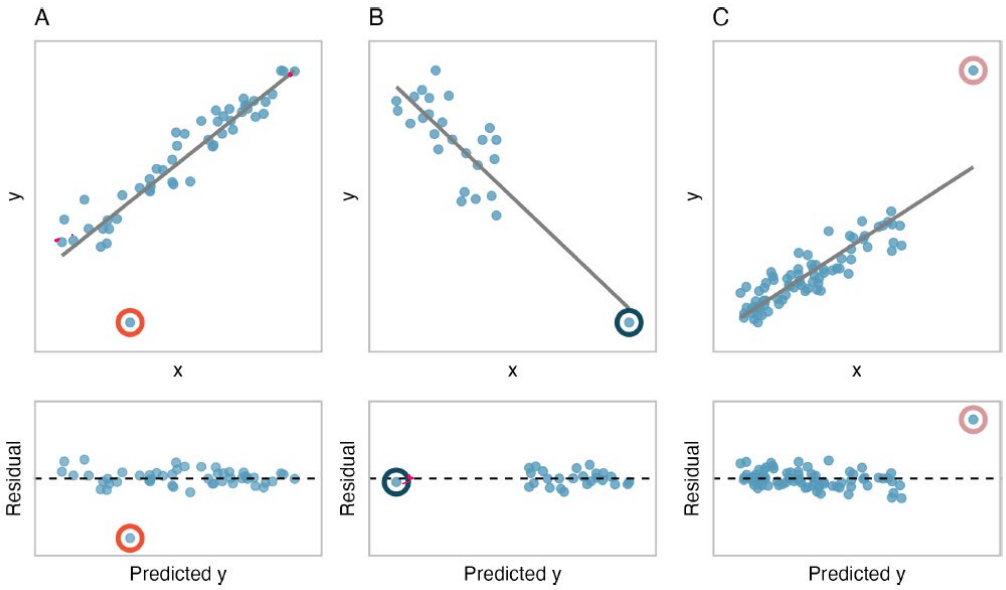


Figure: Three plots, each with a least squares line and corresponding residual plot. Each dataset has at least one outlier.

Leverage

Points that fall horizontally away from the center of the cloud tend to pull harder on the line, so we call them points with **high leverage** or **leverage points**.

Points that fall horizontally far from the line are points of high leverage; these points can strongly influence the slope of the least squares line. If one of these high leverage points does appear to actually invoke its influence on the slope of the line –then we call it an **influential point**. Usually we can say a point is influential if, had we fitted the line without it, the influential point would have been unusually far from the least squares line.

Homework Chapter 7

1. Review section 7.4 (the chapter review)
2. **Suggested:** problems from textbook section 7.5 exercises: 2, 3, 4, 7, 9, 19, 23, 26
3. **Suggested:** begin working on Unit 3 Tutorial Regression Modeling:
   https://openintrostat.github.io/ims-tutorials/03-model/

   1 - Visualizing two variables

   2 - Correlation

   3 - Simple linear regression

   4 - Interpreting regression models

   5 - Model fit

   6 - Parallel slopes