Discussion 13

Hanying Jiang

Week 13 discussion

Announcement: as Jana mentioned on Tuesday, the midterm 2 grades will be adjusted for misconduct soon.

Review

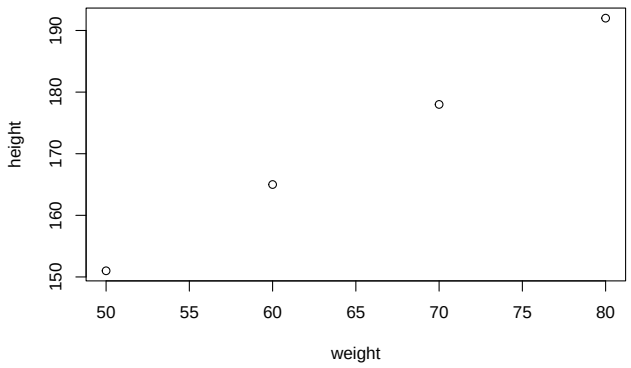
- Notations
- Compute coefficient estimate
- Hypothesis testing of $\hat{\beta}_1$.

Exercise 1

Question

We measured the heights and weights of 4 randomly chosen people between 18 and 24 years old. A plot of the data is given below.

```
weight <- c(50, 60, 70, 80)
height <- c(151, 165, 178, 192)
plot(weight, height)
```



a. Describe the relationship between height and weight. Based on this graph, does a straight-line model seem reasonable?

b. Below is a summary of a linear model fit in R. How do these summary values match the visual interpretation of the data in (a)?

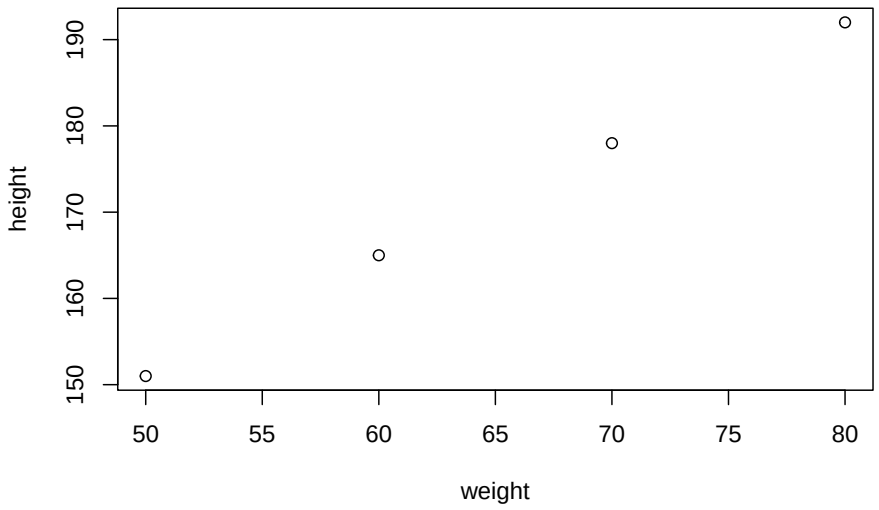
```
disc14_mod <- lm(height ~ weight)
summary(disc14_mod)
```

Solution

(a)

Describe the relationship between height and weight. Based on this graph, does a straight-line model seem reasonable?

```
weight <- c(50, 60, 70, 80)
height <- c(151, 165, 178, 192)
plot(weight, height)
```



We see a very strong, positive, linear relationship between height and weight. Since the points are roughly on a straight line, we can assume a linear model is reasonable.

(b)

Below is a summary of a linear model fit in R. How do these summary values match the visual interpretation of the data in (a)?

```
disc14_mod <- lm(height ~ weight)
summary(disc14_mod)
```

```
##
## Call:
## lm(formula = height ~ weight)
##
## Residuals:
##    1    2    3    4
## -0.1  0.3 -0.3  0.1
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 83.10000    0.93274   89.09 0.000126 ***
## weight       1.36000    0.01414   96.17 0.000108 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3162 on 2 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9997
## F-statistic:  9248 on 1 and 2 DF,  p-value: 0.0001081
```

The slope coefficient for weight has a positive value, with a very small p-value. This matches the fact that we noticed a strong positive relationship.

Exercise 2

Question

The values and some summary measures for the height and weight data are given below. You can also use the fact that $\sum_{i=1}^{4}(x-\bar{x})(y-\bar{y}) = 680$.

Student	x: wts (kg)	y: hts (cm)
1	50	151
2	60	165
3	70	178
4	80	192

Variable	Mean	Standard Deviation
Weight (x)	65	12.91
Height (y)	171.5	17.56

```
weight <- c(50, 60, 70, 80)
height <- c(151, 165, 178, 192)
W.mean <- mean(weight); W.sd <- sd(weight)
H.mean <- mean(height); H.sd <- sd(height)

sum((weight - W.mean)*(height - H.mean)) #680
```

```
## [1] 680
```

For all of the parts below, please perform the calculations "by hand" and show your work. Use built-in R functions and/or the summary in 1(b) to confirm your answers.

In practice, we would typically want way more than 4 points to ensure a linear model is reasonable in the population of points and that the estimates we get for the model are reliable. We have a small sample size in this toy example so the mechanics are easier to learn and practice.

a. Compute the sample correlation and the least square estimates for the y intercept and slope of the regression line relating height (y) to weight (x). Make sure the estimates match the R output in 1(b).

b. Interpret the slope and y intercept estimates from (a) in the context of the question.

c. Write out the least squares regression line and use it to calculate the 4 fitted values of height for the observed weights. Also, calculate the 4 residual values.

d. Create a residual plot (x=fitted values, y=residuals) and a QQ plot of the residuals. Do the necessary regression assumptions seem met?

e. Compute the quantity: $\frac{\sum Residuals^2}{n-2}$. What is this quantity called and what does it mean? Note where this quantity (or a related quantity) can be found in the summary of our lm object.

f. Find the standard error of the estimated slope $\hat{\beta}_1$.

g. Perform a t-test at $\alpha = 0.05$ for the following hypotheses about the slope of the least squares regression line (compute the observed test statistic, p value, and draw a conclusions)

(i)

$$H_0 : \beta_1 \leq 0$$
$$\text{vs.}$$
$$H_A : \beta_1 > 0,$$

(ii)

$$H_0 : \beta_1 = 1$$
$$\text{vs.}$$
$$H_A : \beta_1 \neq 1,$$

Solution

(a)

Compute the sample correlation and the least square estimates for the y intercept and slope of the regression line relating height (y) to weight (x). Make sure the estimates match the R output in 1(b).

```
weight <- c(50, 60, 70, 80)
height <- c(151, 165, 178, 192)
W.mean <- mean(weight); W.sd <- sd(weight)
H.mean <- mean(height); H.sd <- sd(height)
```

```
## W.mean = 65;
## W.sd = 12.9099444873581;
## H.mean = 171.5;
## H.sd = 17.5594229214212.
```

```
## [1] 680
```

Correlation

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{680}{(4-1) * 12.91 * 17.56} = 0.999854.$$

Intercept

$$\hat{\beta}_1 = r(\frac{s_y}{s_x}) = 0.999854 * (\frac{17.56}{12.91}) = 1.36.$$

Slope

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 171.5 - 1.36 * 65 = 83.1.$$

Fitted model

$$\hat{height} = 83.1 + 1.36 * weight.$$

(b)

Interpret the slope and y intercept estimates from (a) in the context of the question.

Intercept 83.1: The average height of a person with no weight is 83.1 cm. This value is not useful for interpretation (it is not good practice to apply your model outside the range of x values were it was built, and a person with weight 0 makes no sense anyway).

Slope 1.36: The average height is expected to increase by 1.36 cm when weight increases by 1 kg.

(c)

Write out the least squares regression line and use it to calculate the 4 fitted values of height for the observed weights. Also, calculate the 4 residual values.

Regression line: $\hat{height} = 83.1 + 1.36 * weight$.

Fitted values are obtained by plugging the x-values into regression equation. Residuals are obtained by subtracting $observed - expected$.

For example, with the observed weight 50 kg, the fitted value is $83.1 + 1.36 * 50 = 151.1$. And the residual is $151 - 151.1 = -0.1$.

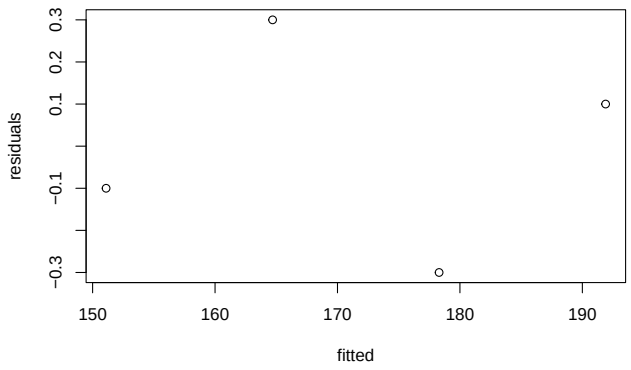
All fitted values are $151.1, 164.7, 178.3, 191.9$. All residuals are $-0.1, 0.3, -0.3, 0.1$.

```
# Compute them by hand
y_hats <- 83.1 + 1.36*weight
resid <- height - y_hats
# Grab them from R
# disc14_mod <- lm(height ~ weight) # recall disc14_mod is the fit object
residuals <- resid(disc14_mod)
fitted <- fitted(disc14_mod)
```

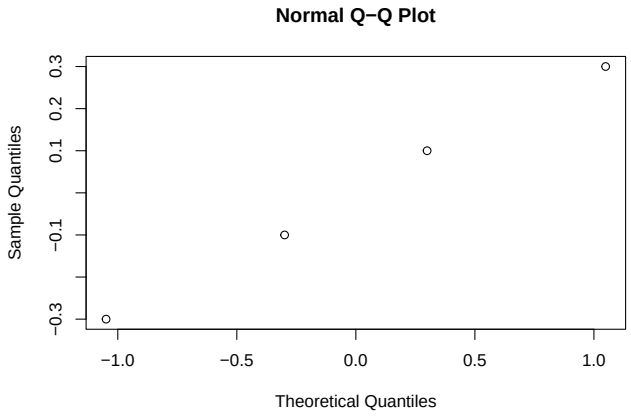
(d)

Create a residual plot (x=fitted values, y=residuals) and a QQ plot of the residuals. Do the necessary regression assumptions seem met?

```
plot(fitted, residuals)
```



```
qqnorm(residuals)
```



Clearly the QQ plot is showing a straight line and hence assumption of normality of errors is fine. The fitted vs residual plot is also showing that the points are randomly scattered and the width around line $y = 0$ across the fitted values are more or less same. We have so few points, however, that it is impossible to check this assumption well from our sample data. The assumption of constant variance is plausible, but ideally we would have more data to evaluate to be more sure.

(e)

Compute the quantity: $\frac{\sum Residuals^2}{n-2}$. What is this quantity called and what does it mean? Note where this quantity (or a related quantity) can be found in the summary of our `lm` object.

$$\frac{\sum Residuals^2}{n-2} = \frac{(-0.1)^2 + 0.3^2 + (-0.3)^2 + 0.1^2}{4-2} = 0.1$$

This quantity is often referred to as Mean Square Error (MSE) and is used to estimate the variance of the population of errors around the least squares regression line.

Residual standard error (σ) is $\sqrt{MSE}$, can be found at the bottom of the summary table.

```
sqrt(0.1)
```

```
## [1] 0.3162278
```

```
summary(disc14_mod)
```

```
##
## Call:
## lm(formula = height ~ weight)
##
## Residuals:
##    1    2    3    4
## -0.1  0.3 -0.3  0.1
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 83.10000    0.93274   89.09 0.000126 ***
## weight       1.36000    0.01414   96.17 0.000108 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3162 on 2 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9997
## F-statistic:  9248 on 1 and 2 DF,  p-value: 0.0001081
```

(f)

Find the standard error of the estimated slope $\hat{\beta}_1$.

$$\hat{se}(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum(x_i - \bar{x})^2}} = \frac{\sigma}{\sqrt{SSx}} = \frac{0.3162278}{\sqrt{3 * 12.91^2}} = 0.01414.$$

(g)

Perform a t-test at $\alpha = 0.05$ for the following hypotheses about the slope of the least squares regression line (compute the observed test statistic, p value, and draw a conclusions)

(i)

$$H_0 : \beta_1 \leq 0$$
$$\text{vs.}$$
$$H_A : \beta_1 > 0,$$

test statistic

$$t_{obs} = \frac{\hat{\beta}_1 - 0}{\hat{se}(\hat{\beta}_1)} = \frac{1.36 - 0}{0.01414} = 96.18105.$$

degrees of freedom

$$df = n - 2 = 4 - 2 = 2.$$

p value

Becaues the alternative hypothesis is ">", the p value is $P(t_2 \geq t_{obs}) = 5.404065e - 05$.

```r
# pt(96.18105, 2, lower.tail = FALSE)
# the code above should give the same results
1 - pt(96.18105, 2)
```

```
## [1] 5.404065e-05
```

Draw conclusion

The p-value is smaller than $\alpha = 0.05$. We reject H_0 and have strong evidence that there is a positive linear relationship between height and weight.

(ii)

$$H_0 : \beta_1 = 1$$
$$\text{vs.}$$
$$H_A : \beta_1 \neq 1,$$

test statistic

$$t_{obs} = \frac{\hat{\beta}_1 - 1}{\hat{se}(\hat{\beta}_1)} = \frac{1.36 - 1}{0.01414} = 25.45969.$$

p value

```
2 * (1 - pt(25.45969, df = 2))
```

```
## [1] 0.001539183
```

Draw conclusion

Since the p-value is very low we can reject $H_0 : \beta_1 = 1$. Evidence suggests the positive linear relationship between height and weight does not have a slope of 1.