Day5

Olivia Wu

2024-02-22

Problem 2.2

False True

Problem 2.3

True

Problem 2.4

False

Problem 2.11

Given: $n = 40, \hat{\beta}_1 = 15.5, SE_{\hat{\beta}_1} = 3.4$, all conditions met

a)

State: $H_0 : \beta_1 = 0 \quad H_a : \beta_1 > 0$

Plan: All conditions met

Do:

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE_{\hat{\beta}_1}} = \frac{15.5 - 0}{3.4} = 4.559$$

$$df = n - 2 = 38$$

$$P(t > 4.559) = 2.61 \times 10^{-}5$$

Conclude: Since $p = 2.61 \times 10^{-5} < 0.05$, we reject the null hypothesis. There is enough evidence to suggest that $\beta_1 > 0$

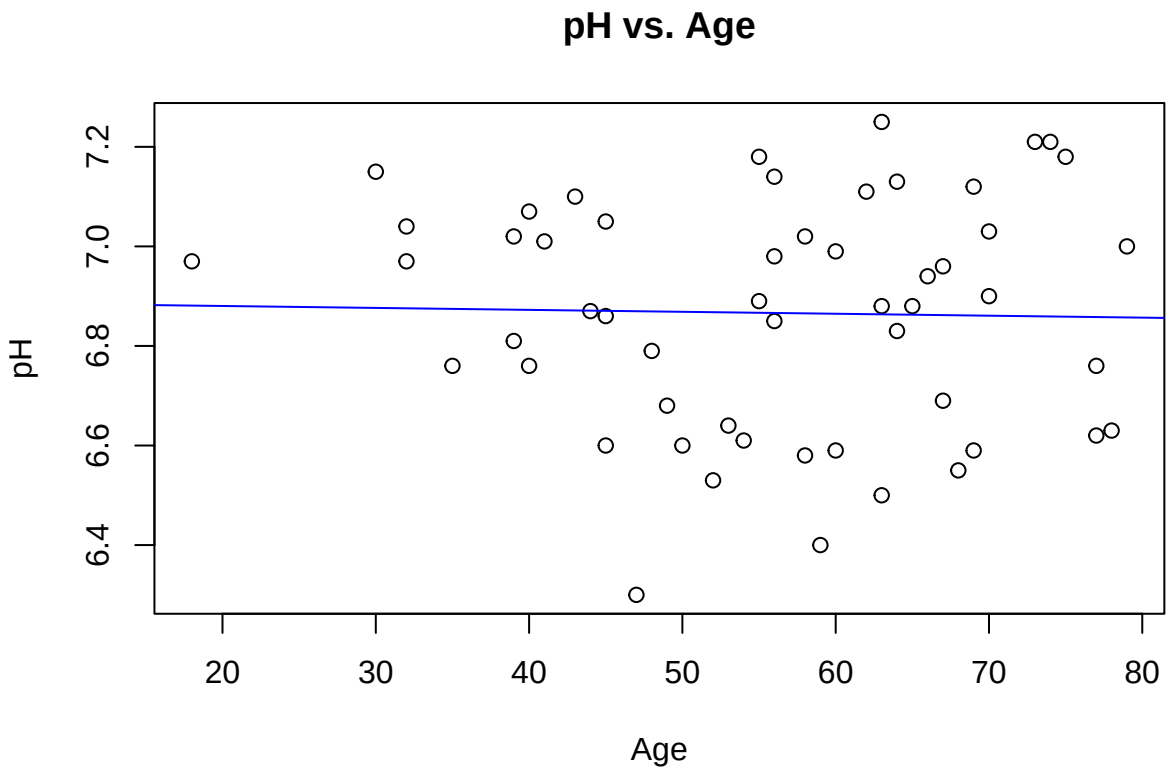
b) We are 95% confident that the true slope of the population regression line lies in the interval (8.617, 22.383).

```r
critT <- qt(0.975, df=38)
left <- sampleBeta -critT*SEbeta
right <- sampleBeta + critT*SEbeta
paste("(",left,",",right,")")
```

```
## [1] "( 8.61705984269931 , 22.3829401573007 )"
```

Problem 2.13

a) There seems to be a very weak negative linear relationship between Age and brain pH.




```
##
## Call:
## lm(formula = pH ~ Age, data = brainpH)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56976 -0.21781  0.02032  0.16801  0.38649
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.8881113  0.1321194   52.13   <2e-16 ***
## Age         -0.0003905  0.0022944   -0.17    0.866
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.235 on 52 degrees of freedom
## Multiple R-squared:  0.0005566,  Adjusted R-squared:  -0.01866
## F-statistic: 0.02896 on 1 and 52 DF,  p-value: 0.8655
```

b) We set up the hypotheses: $H_0 : \beta_1 = 0$ and $H_a : \beta_1 < 0$. The summary output shows the slope has a $t-$value of -0.17 and the p-value $= 0.866 > 0.05$. There is not enough evidence to reject the null hypothesis, which supports our suspicions from part (a).

Problem 2.15

a) $H_0 : \beta_1 = 0 \qquad H_a : \beta_1 > 0$

From the computer output, we know $t = 3.507$ and $p = 0.0013$ for $df = 34$. Since $p < 0.05$, we reject the null hypothesis and there is enough evidence to suggest a linear relationship between sugar content and calories.

```
model <- lm(Calories~Sugar, data=cereal)
summary(model)
```

```
##
## Call:
## lm(formula = Calories ~ Sugar, data = cereal)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.428  -9.832   0.245   8.909  40.322
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  87.4277     5.1627  16.935   <2e-16 ***
## Sugar         2.4808     0.7074   3.507   0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.27 on 34 degrees of freedom
## Multiple R-squared:  0.2656, Adjusted R-squared:  0.244
## F-statistic:  12.3 on 1 and 34 DF,  p-value: 0.001296
```

b) We are 95% confident that the true average increase in calories per gram of sugar lies in the interval (1.04, 3.92).

```
critT <- qt(0.975, df=34)
sampleBeta <- model$coefficients[2]
SEbeta <- 0.7074
left <- sampleBeta - critT*SEbeta
right <- sampleBeta + critT*SEbeta
paste("(",left,",",right,")")
```

```
## [1] "( 1.0432028346347 , 3.91842236641741 )"
```

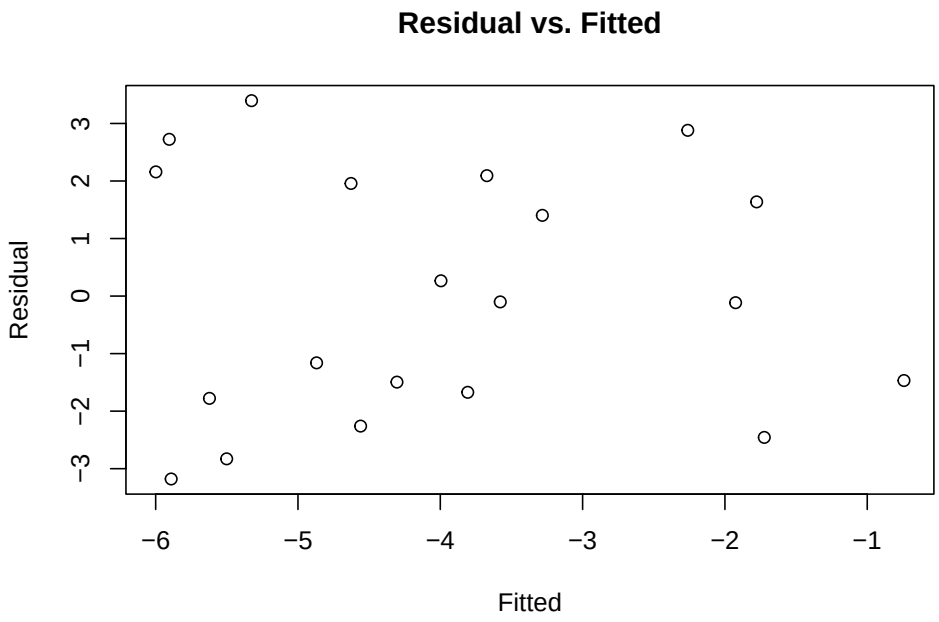
Problem 2.17

a) The p-value of the slope statistic is $0.00516 < 0.05$, so the slope is statistically significant.

```
model <- lm(MMSE~APC, data=LDLB)
summary(model)
```

```
##
## Call:
## lm(formula = MMSE ~ APC, data = LDLB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1791 -1.6991 -0.1081  1.9911  3.3963
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.4359     0.6858  -3.552  0.00228 **
## APC           1.3444     0.4225   3.182  0.00516 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.184 on 18 degrees of freedom
## Multiple R-squared:   0.36,  Adjusted R-squared:  0.3245
## F-statistic: 10.13 on 1 and 18 DF,  p-value: 0.005161
```

b) The residuals are randomly scattered and there is uniform variance, the residuals are centered roughly around zero.



c) $\hat{\beta}_1 = 1.3444, SE_{\beta_1} = 0.4225$

```
summary(model)
```

```
##
## Call:
## lm(formula = MMSE ~ APC, data = LDLB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1791 -1.6991 -0.1081  1.9911  3.3963
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.4359     0.6858  -3.552  0.00228 **
## APC           1.3444     0.4225   3.182  0.00516 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.184 on 18 degrees of freedom
## Multiple R-squared:   0.36,  Adjusted R-squared:  0.3245
## F-statistic: 10.13 on 1 and 18 DF,  p-value: 0.005161
```

d) The confidence interval (0.612, 2.077) does not contain 0, so we can say with 90% confidence there is a statistically significant linear relationship between $MMSE$ and APC.

```
critT <- qt(0.95, df=18)
SEbeta <- 0.4225
sampleBeta <- 1.3444
left <- sampleBeta - critT*SEbeta
right <- sampleBeta + critT*SEbeta
paste("(",left,",",right,")")
```

```
## [1] "( 0.61175812620409 , 2.07704187379591 )"
```

Problem 2.19

a) On average, the 2007 selling price adjusted to 2014 dollars is expected to decrease by \$54.43.

$$\hat{adj2007} = 388.2 - 54.43(distance)$$

```
model <- lm(adj2007~distance,data=RT)
summary(model)
```

```
##
## Call:
## lm(formula = adj2007 ~ distance, data = RT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -190.55  -58.19  -17.48   25.22  444.41
##
```

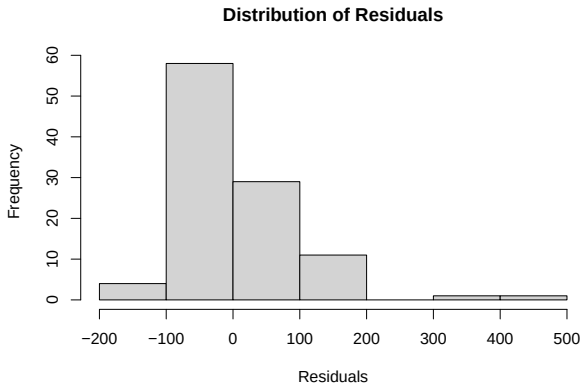
```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  388.204     14.052  27.626  < 2e-16 ***
## distance     -54.427      9.659  -5.635 1.56e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 92.13 on 102 degrees of freedom
## Multiple R-squared:  0.2374, Adjusted R-squared:  0.2299
## F-statistic: 31.75 on 1 and 102 DF,  p-value: 1.562e-07
```

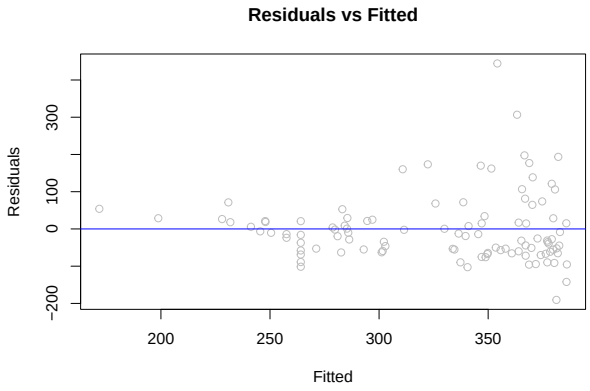
b) We are 90% confident that the true average 2014 dollar increase in 2007 selling price for every foot farther a home is away to a bike trail lies between -$70,460 and -$38,390. write in a better way, dollar increase for every foot CLOSER

```
critT <- qt(0.95, df=102)
sampleBeta <- model$coefficients[2]
SEbeta <- 9.659
left <- sampleBeta - critT*SEbeta
right <- sampleBeta + critT*SEbeta
paste("(",left,",",right,")")
```

```
## [1] "( -70.4604661246307 , -38.393938848836 )"
```

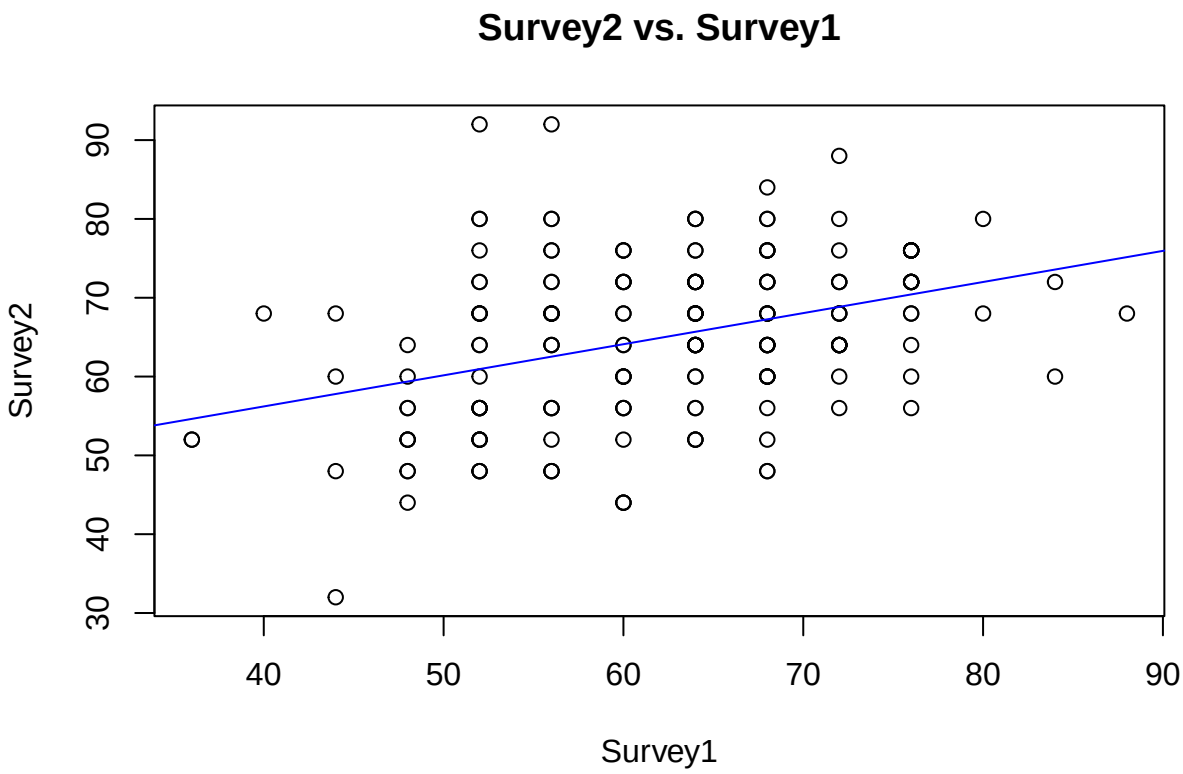
c) The residuals display a strong right skew in their histogram, so the normality condition is not met. The residuals also do not have constant variance, as they grow larger as the fitted values increased. These may affect the accuracy of our confidence interval.





Problem 2.51

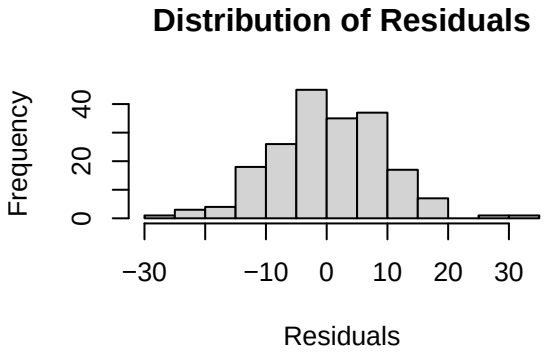
a) There is an overall positive linear relationship between an individual's Survey1 and Survey2 results. A person who does well on Survey1 tends to also do well on Survey2.

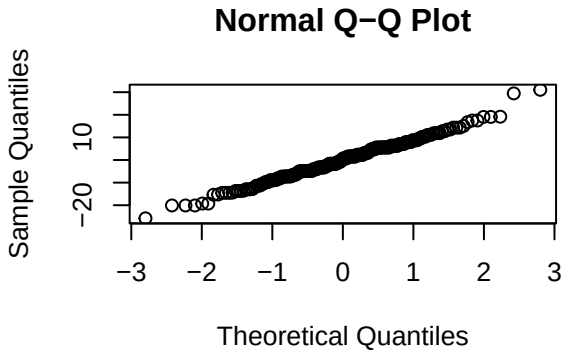


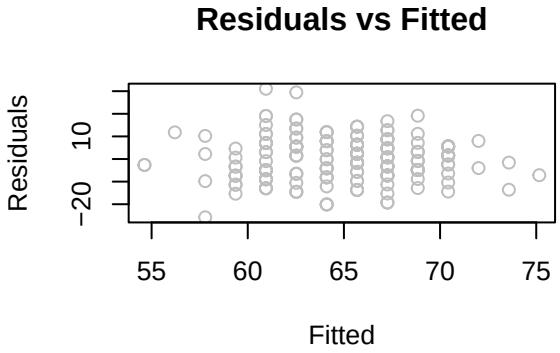
b) $\widehat{Survey2} = 40.417 + 0.395 Survey1$

```
##
## Call:
## lm(formula = Survey2 ~ Survey1, data = Handwriting)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.7870  -6.4722   0.6339   5.9487  31.0547
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.41661    4.43100   9.121  < 2e-16 ***
## Survey1      0.39478    0.07004   5.637 6.07e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.338 on 193 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.1414, Adjusted R-squared:  0.1369
## F-statistic: 31.77 on 1 and 193 DF,  p-value: 6.074e-08
```

c) The distribution of residuals is roughly symmetric. The normal quantile plot of the residuals are roughly linear. There is a uniform variance. This is a randomized experiment, so independence is given.







d) If $Survey1 = Survey2$, we would expect $\beta_0 = 0$ and $\beta_1 = 1$. The computer output shows significant p-values for the regression coefficients, so we know the slope is greater than 0. The standard error for slope is only 0.07, so 1 is out of the question.

Problem 2.52

```
##
## Call:
## lm(formula = SalePrice ~ ListPrice, data = GHouse)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -55942  -3275    846   4141  44168
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.448e+02  5.236e+02  -0.277    0.782
## ListPrice    9.431e-01  3.201e-03 294.578   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8019 on 927 degrees of freedom
## Multiple R-squared:  0.9894, Adjusted R-squared:  0.9894
## F-statistic: 8.678e+04 on 1 and 927 DF,  p-value: < 2.2e-16
```

a) We are 95% confident that the true slope of the population regression line lies between (0.9368, 0.9493).

b) The computer output shows that the p-value for the intercept is $0.782 > 0.05$. Thus, we cannot reject the hypothesis that the true intercept is 0.

c) The confidence interval here is (0.9312, 0.9401). It is less than the interval from (a), where the intercept was negative. After moving it to 0, the slope should be more flat.

```
fraction <- GHouse$SalePrice/GHouse$ListPrice
t.test(fraction, alternative="two.sided", mu=0)
```

```
##
##  One Sample t-test
##
## data:  fraction
## t = 410.11, df = 928, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.9311611 0.9401158
## sample estimates:
## mean of x
## 0.9356384
```

Problem 2.61

a)

$$\hat{\beta}_1 = 0.701\frac{104807}{657} = 111.826\hat{\beta}_0 = 247235 - \hat{\beta}_1 2009 = 22576.566\hat{Gate} = 22576.556 + 11.826Enroll$$

b) 49.14% of the variation is accounted for by the enrollments.

$$r^2 = 0.701^2 = 0.4914$$

c)

$$\hat{Gate} = 22576.566 + 111.826 \cdot 1445 = \boxed{184165 \text{ people}}$$

d)

$$\hat{Gate} = 22576.566 + 111.826 \cdot 2200 = 268593.766\text{residual} = 130,000 - \hat{Gate} = \boxed{-138593.766}$$