Homework 5

Joshua Ingram

3/6/2020

Problem 1

1.

a.

$education_i = \alpha + \beta_1 income_i + \epsilon_i, \epsilon_i \overset{i.i.d}{} N(0, \sigma^2)$

b.

Hypothesis Test:

$H_0 : \beta_1 = 0$

$H_A : \beta_1 \neq 0$

$\alpha = 0.05$

c.

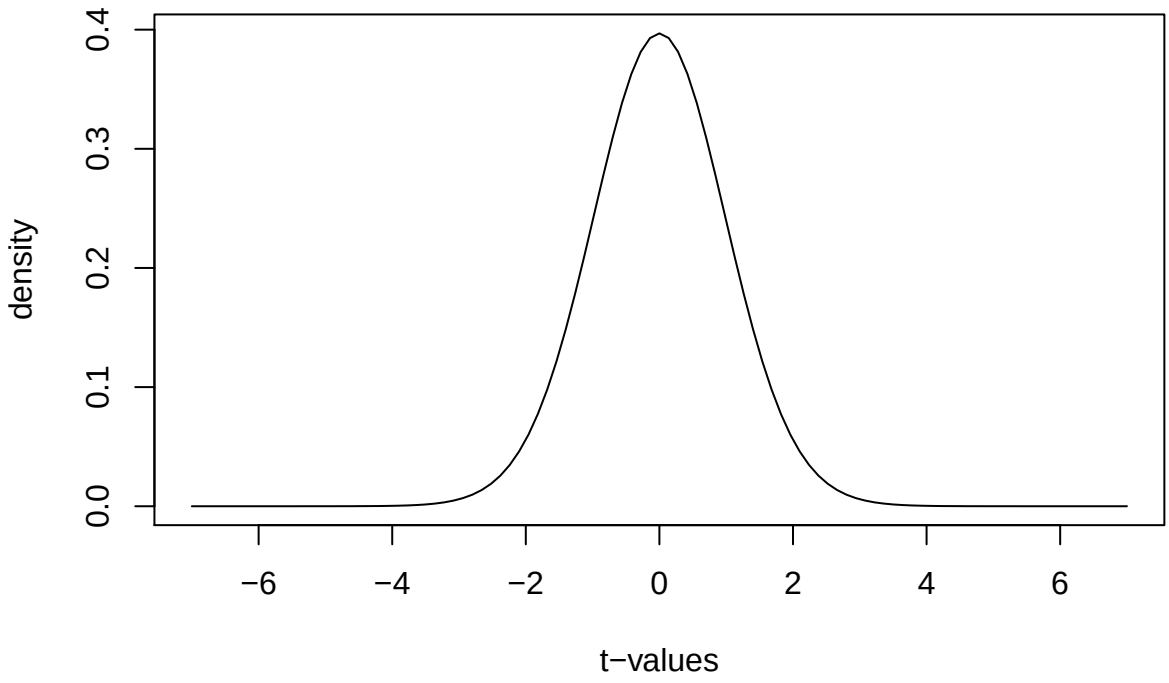
```
lm_edu <- lm(education ~ income, data = anscombe)
summary_1 <- summary(lm_edu)
t_value <- summary_1$coefficients[2, 3]
```

With a t-value of 6.276 and p-value of $8.76e^{-8}$ (very small), we have significant evidence to reject the null hypothesis in favor of the alternative, that there is a relationship between education spending and income.

d.

```
df_1 <- nrow(anscombe) - 2
alpha <- 0.05

# creating the t-distribution with 2 degrees of freedom
curve(dt(x, df_1), -7, 7, ylab = "density", xlab = "t-values")
```




```
# p-value calculation
pt(t_value, df = df_1, lower.tail = F) * 2
```

```
## [1] 8.762267e-08
```

The p-value for our hypothesis is found by calculating the t-value, then using the t-distribution with 2 degrees of freedom (for this specifc case) to find the sum of the area under the curve to the right of 6.28641 and to the left of -6.28641 (since this is a two-tailed test). This will give us our p-value. The graph above is the t-distribution with 2 degrees of freedom, so we would be using this distribution to find our p-value.

Using the "pt()" function and the t-value output from our summary, I found the same p-value of $8.762e^{-8}$.

e.

```
confint(lm_edu, level = .99)
```

```
##                    0.5 %     99.5 %
## (Intercept) -59.67047466 95.0905362
## income        0.03173108  0.0790208
```

We are 99% confident that, on average, for every dollar increase in per-capita income, we will see between a 0.03173 and 0.07902 dollar increase in education expenditures per-capita.

2.

a.

```
lm_edu2 <- lm(education ~ income + under18, data = anscombe)
summary_2 <- summary(lm_edu2)
summary_2$coefficients
```

```
##                  Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) -301.08922003 70.271344297 -4.284666 8.749701e-05
## income         0.06118383  0.007413335  8.253213 9.149047e-11
## under18        0.83610633  0.173274822  4.825319 1.457609e-05
```

```
confint(lm_edu2, level = 0.95)
```

```
##                     2.5 %        97.5 %
## (Intercept) -442.37922734 -159.79921272
## income         0.04627832    0.07608934
## under18        0.48771395    1.18449871
```

Income $H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

$\alpha = 0.05$

With a t-value of 8.2532 and p-value of $9.149e^{-11}$, we have significant evidence to reject the null hypothesis in favor of the alternative, that there is a relationship between income and education spending.

We are 95% confident that, on average, for every dollar increase in per-capita income, we will see between a 0.0463 and 0.0761 dollar increase in education expenditures per-capita.

Under18 $H_0 : \beta_2 = 0$

$H_a : \beta_2 \neq 0$

$\alpha = 0.05$

With a t-value of 4.825 and p-value of $1.4576e^{-5}$, we have significant evidence to reject the null hypothesis in favor of the alternative, that there is a relationship between the proportion under 18 and education spending.

We are 95% confident that, on average, for every 1 unit increase in the proportion under 18, we will see between a 0.4877 and 1.1845 dollar increase in education expenditures per-capita.

b.

```
lm_edu3 <- lm(education ~ income + under18 + urban, data = anscombe)
summary_3 <- summary(lm_edu3)
summary_3$coefficients
```

```
##                  Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) -286.83876273 64.919931523 -4.418347 5.823463e-05
## income         0.08065325  0.009298538  8.673756 2.563747e-11
## under18        0.81733774  0.159789699  5.115084 5.694503e-06
## urban         -0.10580623  0.034282173 -3.086334 3.392812e-03
```

Income $H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

$\alpha = 0.05$

With a t-value of 8.674 and p-value of $2.563747e^{-11}$, we have significant evidence to reject the null hypothesis in favor of the alternative, that there is a relationship between income and education spending.

Under18 $H_0 : \beta_2 = 0$

$H_a : \beta_2 \neq 0$

$\alpha = 0.05$

With a t-value of 5.115084 and p-value of $5.694503e^{-6}$, we have significant evidence to reject the null hypothesis in favor of the alternative, that there is a relationship between the proportion under 18 and education spending.

Urban $H_0 : \beta_3 = 0$

$H_a : \beta_3 \neq 0$

$\alpha = 0.05$

With a t-value of -3.086334 and p-value of $3.392812e^{-3}$, we have significant evidence to reject the null hypothesis in favor of the alternative, that there is a relationship between the proportion urban and education spending.

c.

Standard Error from part (a): 0.007413335

Standard Error from part (b): 0.009298538

The standard error from part (b), where we added a third variable, is larger. This means our estimate of β_1 is less stable. This increase is likely due to some collinearity between income and under18/urban

```
lm_inc1 <- lm(income ~ under18, data = anscombe)
lm_inc2 <- lm(income ~ under18 + urban, data = anscombe)

summary(lm_inc1)$r.squared
```

```
## [1] 0.02636076
```

```
summary(lm_inc2)$r.squared
```

```
## [1] 0.4744752
```

The R^2 for the model regressing income onto under18 is noticeably smaller (0.0263) than R^2 for the model regressing income onto under18 and urban (.4745). This increase in R^2 explains the increase in the standard error of $\hat{\beta}_1$, as the formula for the variance of an estimate $\hat{\beta}_j$ takes into the the R^2 value in the VIF. So as R^2 increases, the variance of $\hat{\beta}_j$ will increase and consequently its standard error. This explains the decrease in stability of our estimate when we added urban, because there is collinearity between the predictors in the model in part (b).

Problem 2

1.

a.

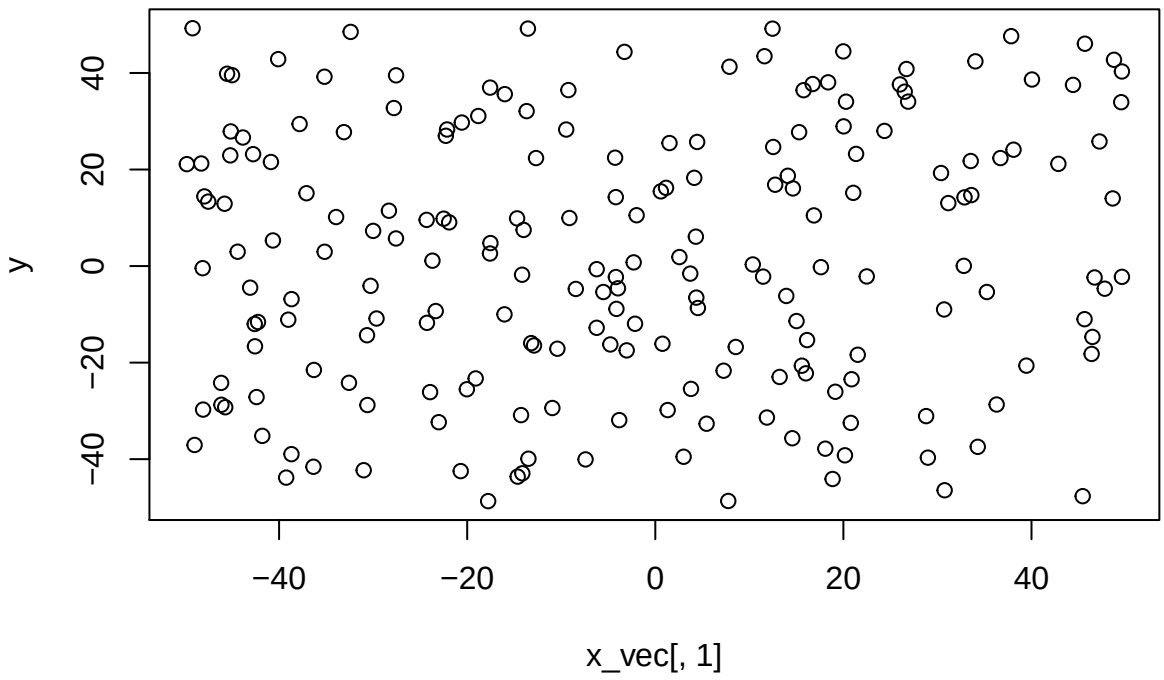
```
set.seed(50)
x_vec <- matrix(runif(20000, -50,50), ncol=100)
```

b.

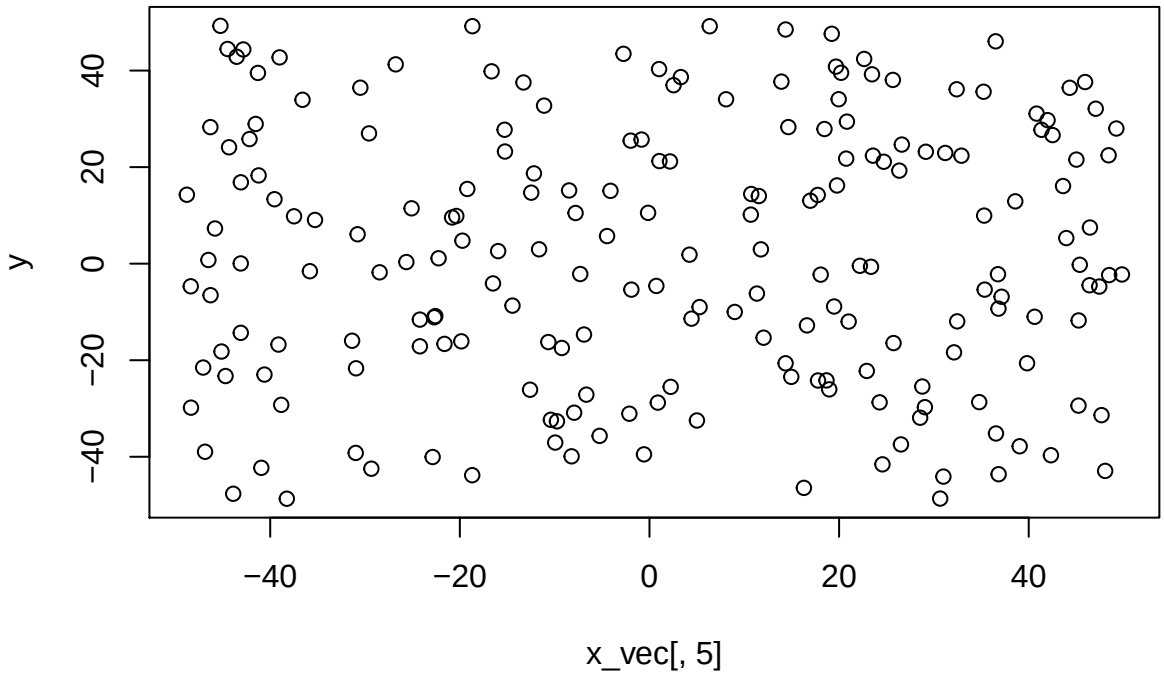
```
set.seed(1)
y <- runif(200, -50, 50)
```

c.

```
plot(y ~ x_vec[,1])
```



```r
plot(y ~ x_vec[,5])
```



2.

a.

$H_0 : \beta_1 = \beta_2 = ... = \beta_{100} = 0$

$H_a : \exists \text{ j} \mid \beta_j \neq 0$

$\alpha = 0.05$

```r
lm_y <- lm(y ~ x_vec)
summary_y <- summary(lm_y)
summary_y$fstatistic
```

```
##       value       numdf       dendf
##   0.8011889 100.0000000  99.0000000
```

```r
pf(summary_y$fstatistic[1], 100, 99)
```

```
##     value
## 0.1351988
```

With an f-statistic of 0.8011880 and p-value of 0.1351988, we do not have significant evidence to reject the null hypothesis, that there is no relationship between any of our predictors and y.

b.

```
sum(summary_y$coefficients[2:100,4] <= 0.05)
```

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

There was 1 signifcant p-value amongst all the t-tests (exluding the intercept).

c.

This single "significant" relationship is due to pure chance, as we know that our y-value was determined completely randomly and was not a function of our predictors at all.

d.

This would be a type 1 error, as we rejected true null hypothesis.

Problem 3

1.

```
f_test <- function(y, x){
  lm <- lm(y ~ x)
  n <- length(y)
  df_num <- ncol(x)
  df_denom <- n - (df_num + 1)
  rss <- sum(lm$residuals^2)
  tss <- sum((y - mean(y))^2)
  regss <- tss - rss
  f <- (regss/df_num)/(rss/df_denom)
  p_val <- pf(f, df_num, df_denom, lower.tail = F)
  list <- list("F" = f, "p.val" = p_val)
  return(list)
}
```

2.

```
predictors <- as.matrix(data.frame(anscombe$income, anscombe$under18, anscombe$urban))
f_test(anscombe$education, predictors)
```

```
## $F
## [1] 34.81053
##
## $p.val
## [1] 5.33677e-12
```

a.

$H_0 : \beta_1 = \beta + 2 = \beta_3 = 0$

$H_a : \exists \text{ j} \mid \beta_j \neq 0$

$\alpha = 0.05$

b.

With a f-statistic of 34.81053 and a p-value of $5.33677e^{-11}$, we have significant evidence to reject the null hypothesis in favor of the alternative, that there is a relationship between education spending and at least one of our predictors.

3.

```
summary(lm(education ~ income + under18 + urban, data = anscombe))
```

```
##
## Call:
## lm(formula = education ~ income + under18 + urban, data = anscombe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.240 -15.738  -1.156  15.883  51.380
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.868e+02  6.492e+01  -4.418 5.82e-05 ***
## income       8.065e-02  9.299e-03   8.674 2.56e-11 ***
## under18      8.173e-01  1.598e-01   5.115 5.69e-06 ***
## urban       -1.058e-01  3.428e-02  -3.086  0.00339 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.69 on 47 degrees of freedom
## Multiple R-squared:  0.6896, Adjusted R-squared:  0.6698
## F-statistic: 34.81 on 3 and 47 DF,  p-value: 5.337e-12
```

The f-statistics are the same with the p-values being the same.