Homework 4

Miles Tweed

2/28/2021

Problem 1

Part 1.a

```
est.beta <- function(X,Y){
  ex.val.x <- mean(X)
  ex.val.y <- mean(Y)
  dif.x <- X - ex.val.x
  dif.y <- Y - ex.val.y
  sq.dif.x <- dif.x**2
  numer <- sum(dif.x * dif.y)
  denom <- sum(sq.dif.x)
  beta1 <- numer/denom
  beta0 <- ex.val.y - beta1*ex.val.x

  return(c(beta0,beta1))
}
```

Part 1.b My estimation of beta0 and beta1 resulted in the same numbers returned using lm().

```
ads <- read_csv('../../Data/Advertising.csv')

lm.obj <- lm(sales~TV, data = ads)

X <- ads$TV
Y <- ads$sales

# Coefficients (beta0, beta1)
lm.obj$coefficients
```

```
## (Intercept)          TV
##  7.03259355  0.04753664
```

```
est.beta(X,Y)
```

```
## [1] 7.03259355 0.04753664
```

Part 2.a

```
err.stat.calc <- function(X,Y){
  beta <- est.beta(X,Y)
  beta0 <- beta[1]
  beta1 <- beta[2]
  y.hat <- beta0 + beta1 * X
  n <- length(Y)
```

```
  y.bar <- mean(Y)
  rse <- sqrt((1/(n-2))*sum((Y-y.hat)**2))
  r.sq <- (sum((Y-y.bar)**2)-sum((Y-y.hat)**2))/sum((Y-y.bar)**2)
  return(c(rse,r.sq))
}
```

Part 2.b

The estimation using my custom function resulted in the same numbers for RSE and R^2.

```
stats <- err.stat.calc(X,Y)

# RSE and R^2
summary(lm.obj)
```

```
##
## Call:
## lm(formula = sales ~ TV, data = ads)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## TV          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
print(paste("RSE=",round(stats[1],3),"R^2=",round(stats[2],4)))
```

```
## [1] "RSE= 3.259 R^2= 0.6119"
```

Problem 2

Show that:

$$Y_i = \beta_0 + \epsilon_i, \epsilon_i \sim_{i.i.d} \mathbb{N}(0, \sigma^2)$$

Leads to:

$$Y_i \sim_{i.i.d} \mathbb{N}(\beta_0, \sigma^2)$$

Here, β_0 is a constant predictor of the response Y as well as the intercept and ϵ_i is a variable that represents the random error associated with the i^{th} observation. The prediction error for one observation (ϵ_i) is considered independent of the others (for example ϵ_{i+1}), since it is assumed that the error for the estimate of Y_i does not effect the error in other estimates (for example Y_{i+1}) because it is random. If β_0 represents the constant line of best fit, it is defined in such a way that there is an approximately equal amount of error above and below the line so the mean of ϵ is assumed to be zero. In other words, the mean is zero because you do not want to have a predictor that is biased towards over- or under-estimates. By definition, the error represents the variation in Y around the estimate β_0 and so has a variance of σ^2 leading to its distribution as normal in accordance with the central limit theorem $\epsilon_i \sim \mathbb{N}(0, \sigma^2)$.

a

We need to calculate the expectation value of both sides of the original equation (1). We can separate and take the expectation values of the two terms independently (2). Since β_0 is a constant its expectation value is simply the constant value (3) and as discussed above, if β_0 is an unbiased estimate of Y, the expectation value (or mean) of epsilon is zero (4). Therefore, the expectaion value of Y_i is β_0 (5).

$$\begin{aligned} \mathrm{E}[Y_i] &= \mathrm{E}[\beta_0 + \epsilon_i] && (1)\\ &= \mathrm{E}[\beta_0] + \mathrm{E}[\epsilon_i] && (2)\\ &= \beta_0 + \mathrm{E}[\epsilon_i] && (3)\\ &= \beta_0 + 0 && (4)\\ &= \beta_0 && (5) \end{aligned}$$

b

Taking the variance of both sides of the original equation (1) and considering the variance of each term separately (2), the variance of a constant is zero because a constant does not change for any value of x (3). As established above, ϵ is normally distributed with a mean of 0 and a standard deviation of σ. Therefore, it's variance is σ^2 (4).

$$\begin{aligned} \mathrm{V}[Y_i] &= \mathrm{V}[\beta_0 + \epsilon_i] && (6)\\ &= \mathrm{V}[\beta_0] + \mathrm{V}[\epsilon_i] && (7)\\ &= 0 + \mathrm{V}[\epsilon_i] && (8)\\ &= \sigma^2 && (9) \end{aligned}$$

c

We defined Y_i in the original equation as $Y_i = \beta_0 + \epsilon_i$ and we established that β_0 was a constant value with an expected value of β_0 and a variance of 0 while ϵ was normally distributed with a mean of 0 and variance σ. The addition of a constant to every point of a normal distribution will not change its shape since the same amount is added to every point, it will not change its variance since the variance of a constant is 0, but it will shift its mean by the constant amount. Since we have defined Y_i as the sum of a constant and a normally distributed error term, we should expect Y_i to be normally distributed with a mean of β_0 and variance σ^2. Or:

$$Y_i \sim \mathbb{N}(\beta_0, \sigma^2)$$

Problem 3

See the Lab_MT.pdf