Assignment 2

Sohaib Syed

2022-09-17

Contents


```
library(magick)
```

```
## Linking to ImageMagick 6.9.12.3
## Enabled features: cairo, freetype, fftw, ghostscript, heic, lcms, pango, raw, rsvg, webp
## Disabled features: fontconfig, x11
```

Problem 1

a

Alternatives

Null Hypothesis: Beta_1=0; Alternative Hypothesis: Beta_1 < 0

Decision Rule

Since we are testing that Beta_1=0, we can use T_obs=BetaHat_1/(simgaHat/squareroot(S_xx)) Also, if testing Beta_1=0 is equal to testing rho=0, then Null hypothesis Beta_1=0 is under t-distribution with n-2 degrees of freedom. Thus to reject the null hypothesis: |T_obs| >= t_(1-alpha,n-2)

```
muscle_data<-read.delim('CH01PR27.txt',sep="",header=FALSE)
colnames(muscle_data)<- c("Y", "X")
muscle_fitted<-lm(Y~X,data=muscle_data)
summary(muscle_fitted)
```

```
##
## Call:
## lm(formula = Y ~ X, data = muscle_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1368  -6.1968  -0.5969   6.7607  23.4731
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 156.3466     5.5123   28.36   <2e-16 ***
## X            -1.1900     0.0902  -13.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.173 on 58 degrees of freedom
## Multiple R-squared:  0.7501, Adjusted R-squared:  0.7458
## F-statistic: 174.1 on 1 and 58 DF,  p-value: < 2.2e-16
```

The summary displays that the T_obs=-13.19, thus the |T_obs|=13.19. Also, since this is a one-sided test, we do not need to use alpha/2 for t-distribution rather we can use simply alpha. The summary also helps us to find that the degrees of freedom for our data set is 58. This gives us that t-distribution with alpha=0.05 and df=58 is equal to 1.67. According to the decision rule 13.19>= 1.67

Conclusion

We reject the null hypothesis because the observed T value was less than value of tvalue in the table. The P-value for this test is 1.1e-16, because of the 1 sided test I took the output from summary and divided that value by 2.

b

No. even though the test shows statistical significance, there wasn't data collected on newborn females. The mass of a newborn does not compare to the mass of an adult. This is why domain knowledge is important in statistics.

c

Beta_1 shows the difference in muscle mass between women whose ages differs by 1 year.

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

```
##                   2.5 %     97.5 %
## (Intercept) 145.312572 167.380556
## X            -1.370545  -1.009446
```

The confident interval at 95% is (-1.370545, -1.009446). It isn't necessary to know the specific equations because the confidence interval equation does not depend on the X or input rather on the slope, t-distribution, and standard error

Problem 2

a

```
gpa_data<-read.delim('CH01PR19.txt',sep="",header=FALSE)
colnames(gpa_data)<- c("Y", "X")
gpa_fitted<-lm(Y~X,data=gpa_data)
Sxx <- sum(gpa_data$X * gpa_data$X) - length(gpa_data$X) * (mean(gpa_data$X))^2
Syy <- sum( gpa_data$Y * gpa_data$Y) - length(gpa_data$Y) * (mean(gpa_data$Y))^2
Sxy <- sum(gpa_data$X *gpa_data$Y ) - length(gpa_data$X) * mean(gpa_data$X) * mean(gpa_data$Y)

beta1hat <- Sxy / Sxx
beta0hat <- mean(gpa_data$Y) - beta1hat * mean(gpa_data$X)
beta1hat
```

```
## [1] 0.03882713
```

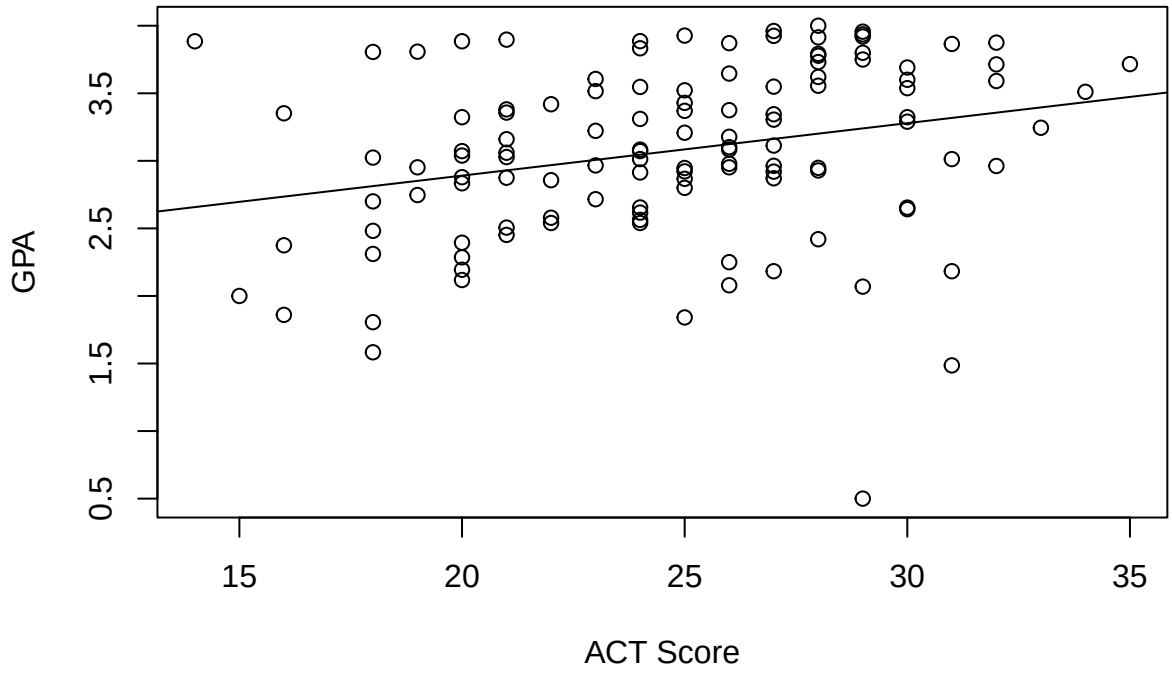
```
beta0hat
```

```
## [1] 2.114049
```

Therefore the estimated regression funciton is Yhat=2.11409+0.03882713x

b

```
plot(gpa_data$X,gpa_data$Y,xlab = 'ACT Score', ylab = 'GPA')
abline(a=beta0hat,b=beta1hat)
```



The estimated regression function fits the data but not too well. The data seems hard to fit with a line.

c

when X=30 with regression line Yhat=2.11409+0.03882713(30)=3.279

d

The point estimate of the change in the mean when test score increases by 1 is the slope which is 0.0388

Problem 3

a

```
xnew <- data.frame(X = c(28))
predict(gpa_fitted, xnew, interval="confidence", level=0.95)
```

```
##        fit      lwr      upr
## 1 3.201209 3.061384 3.341033
```

The interval is (3.061,3.341). This means that linear model is confident that 95% of freshman students with an ACT score of 28 will have a gpa between 3.061 and 3.341

b

```
predict(gpa_fitted, xnew, interval="prediction", level=0.95)
```

```
##        fit      lwr      upr
## 1 3.201209 1.959355 4.443063
```

The prediction interval is (1.959,4.443). This means that with confidence of 95%, the model predicts that with Mary Jones' ACT score of 28, she will have a freshman GPA between 1.959 and 4.443

c

Yes the interval for the prediction is wider than the confidence interval. It makes sense for it to be wider as the value for the prediction interval is random and possibly not learned by the model

d

Taking approach from lecture slides with Dwaine Studio example but only 1 parameter X

```
pred <- predict(gpa_fitted,newdata=data.frame(X=28),se.fit=TRUE)

W <-sqrt(2*qf(1-0.05,2,length(gpa_data$X)-2))
CI_band_upper <- pred$fit+W*pred$se.fit
CI_band_lower <- pred$fit-W*pred$se.fit

CI_band_upper
```

```
##        1
## 3.376258
```

```
CI_band_lower
```

```
##        1
## 3.026159
```

The confidence band is wider than the confidence interval and it makes sense because the confidence band represents confidence intervals for entire regression line.

Problem 4

a

```r
anova(gpa_fitted)
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df Sum Sq Mean Sq F value   Pr(>F)
## X           1  3.588  3.5878  9.2402 0.002917 **
## Residuals 118 45.818  0.3883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
tab <- matrix(nrow=3,ncol=3)

#define column names and row names of matrix
colnames(tab) <- c('SS', 'df', 'MS')
rownames(tab) <- c('Regression', 'Error', 'Total')

#convert matrix to table
tab <- as.table(tab)

#view table
tab[1,1]=3.588
tab[1,2]=1
tab[1,3]=3.5878
tab[2,1]=45.818
tab[2,2]=118
tab[2,3]=.3883
tab[3,1]=tab[1,1]+tab[2,1]
tab[3,2]=119
tab
```

```
##                  SS       df       MS
## Regression   3.5880   1.0000   3.5878
## Error       45.8180 118.0000   0.3883
## Total       49.4060 119.0000
```

b

MSR in the Anova table estimates SSR/(p-1). So for the given data the MSR is 3.58. MSE and MSR estimate the same value when BetaHat_1 is 0.

c

Alternatives

H0:Beta_1 = 0 Ha:Beta_1 != 0

alpha=.01

decision rule

Reject H0 if F-ratio > Fa,p-1,n-p

```
F_ratio<-tab[1,3]/tab[2,3]
F_ratio
```

```
## [1] 9.239763
```

```
F_ratio_from_table<-qf(.99,1,118)
F_ratio_from_table
```

```
## [1] 6.854641
```

```
decision<-F_ratio>F_ratio_from_table
decision
```

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

conclusion

We reject H0 because 9.239763 > 6.854641

d

```
abs_red<-tab[1,1]
abs_red
```

```
## [1] 3.588
```

```
rel_red<-tab[1,1]/tab[3,1]
rel_red
```

```
## [1] 0.07262276
```

The absolute magnitude of the reduction is shown by SSR which the ANOVA table shows as 3.588. The relative reduction takes the absolute reduction of the variance in relation to the total sum of squares. This value is SSR/SStot=0.07262276. This relative reduction is named the Coefficient of Determination.

e

For simple linear regression $r^{2=R}2$, so r= sqaure_root(R^2)

```
sqrt(rel_red)
```

```
## [1] 0.2694861
```

r=+.2694861

f

R^2 has a more clear-cut operational interpretation because it shows the percentage of variation by the linear model as described by its definition.

Problem 5

```
newlogo <- image_read("./hw2p5.jpg")
newlogo <- image_scale(newlogo, "720x1080")
image_rotate(newlogo, 90)
```

5) $t = \frac{\hat{\beta}_1}{s.e(\hat{\beta}_1)}$ $\quad F = \frac{SS_{reg}}{\hat{\sigma}^2}$

$t^2 = \left(\frac{\hat{\beta}_1}{s.e(\hat{\beta}_1)}\right)^2 = \frac{\hat{\beta}_1^2}{s.e(\hat{\beta}_1)^2}$

$SS_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - \hat{\beta}_0 + \hat{\beta}_1 \bar{x})^2$

$= \sum_{i=1}^{n} (\hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x})^2$

$= \sum_{i=1}^{n} (\hat{\beta}_1 (x_i - \bar{x}))^2$

$\hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$

$\hat{\beta}_1^2 S_{xx} = SS_{reg}$

$\frac{\hat{\beta}_1^2}{\frac{\hat{\sigma}^2}{S_{xx}}} = \frac{\hat{\beta}_1^2 S_{xx}}{\hat{\sigma}^2} = \frac{SS_{reg}}{\hat{\sigma}^2}$

$t^2 = F$