STAT512 HW1

Yi Yang

9/4/2018

1. A regresssion analysis relating test scores (Y) to training hours (X) produced the following fitted equation: $\hat{Y} = 20 - 0.6X$. Be sure to show your work for all parts in this question.

a) What is the fitted value of the response variable corresponding to $X = 4$?
b) What is the residual corresponding to the data point with $X = 5$ and $Y = 21$?
c) If X decreases by 3 units, how does $\hat{Y}$ change?
d) An additional test score is to obtained for a new observation at $X = 7$. Would the test score for the new observation necessarily be 15.8? Explain.
e) The error of sums of squares (SSE) for this model was found to be 15.3. If there were $n = 20$ observations, provide the best estimate for σ^2.
f) Rewrite the regression equation in terms of X^* where X^* is training time mesured in minutes. Show that your answer makes sense, i.e., gives the same predictions as the original equation (one example is sufficient).

Solution:

a) The fitted value of the response variabe is $\hat{Y}(X = 4) = 20 - 0.6 \times 4 = 17.6$.
b) The resiual is given by
$$e = Y - \hat{Y} = 21 - (20 - 0.6 \times 5) = 4$$
c) $\hat{Y}$ will increase by 1.8.
d) No, the new test score is a random variable, but 15.8 is the fitted test score obtained from the regression line.
e) The best estimate for σ^2 is given by
$$s^2 = \frac{SSE}{n-2} = \frac{15.3}{18} = 0.85$$
f) Since $X = \frac{X^*}{60}$, we have
$$\hat{Y} = 20 - 0.6\frac{X^*}{60} = 20 - 0.01X^*$$
Consider that $X = 4$ and $X^* = 240$, the predicted response varibale is 17.6, which is the same as the one obtained in a).

2. The director of admission of a small college selected 10 students at random from the new freshman class in a study to determine whether a student's GOA at the end of the freshman year can be predicted from the ACT test score. The data is GPA.csv.

a) Compute a linear regression to predict GPA based on the ACT test score.
b) Compute the residual standard deviation, s.
c) Give a point estimate and a 95% confidence interval for the slope and intercept and interpret each of these in words. (Point estimate is another word for parameter estimate.) Why would someone be interested if 0 is included in the interval for β_1?
d) Perform a hypothesis test on whether ACT test score is associated with GPA.

Solution:

a) Let us compute the linear regression model
$$\hat{Y}_i = b_0 + b_1 X_i$$

```r
file_df <- read.csv('data/GPA.csv', header = TRUE)
names(file_df)
```

```
## [1] "GPA" "ACT"
```

```r
GPA_score <- file_df$GPA
ACT_score <- file_df$ACT
SSxy <- t((GPA_score - mean(GPA_score))) %*% (ACT_score - mean(ACT_score))
SSx <- t((ACT_score - mean(ACT_score))) %*% (ACT_score - mean(ACT_score))
b1 <- SSxy / SSx
b0 <- mean(GPA_score) - b1 * mean(ACT_score)
print(paste('The slope is', b1))
```

```
## [1] "The slope is 0.0101465428276574"
```

```r
print(paste('The intercept is', b0))
```

```
## [1] "The intercept is 2.90825799793602"
```

```r
print(paste('The predictor model is given by GPA = ', b0, '+', b1, 'ACT'))
```

```
## [1] "The predictor model is given by GPA =  2.90825799793602 + 0.0101465428276574 ACT"
```

b) The residual standard deviation is given by

$$s = \sqrt{\frac{SSE}{n-2}}$$

```r
GPA_pd <- b0 + b1 * ACT_score
GPA_residual <- GPA_score - GPA_pd
SSE <- sum(GPA_residual^2)
n <- length(GPA_score)
s <- sqrt(SSE / (n - 2))
print(paste('The residual standard deviation is', s))
```

```
## [1] "The residual standard deviation is 0.229899847512957"
```

c) A point estimate for slope β_1 is b_1, and a point estimate for intercept β_0 is b_0.

```r
sb1 <- s / sqrt(SSx) # this is the unbiased estimator for sigma{b1}
CI_lb <- b1 - sb1 * qt(1 - 0.05/2,n-2)
CI_ub <- b1 + sb1 * qt(1 - 0.05/2,n-2)
print(paste('The confidence interval for the slope is [', CI_lb, ',', CI_ub, '].'))
```

```
## [1] "The confidence interval for the slope is [ -0.0437098021813786 , 0.0640028878366933 ]."
```

```r
sb0 <- s * sqrt(1 / n + (mean(ACT_score))^2 / SSx)
CI_lb0 <- b0 - sb0 * qt(1 - 0.05/2,n-2)
CI_ub0 <- b0 + sb0 * qt(1 - 0.05/2,n-2)
print(paste('The confidence interval for the intercept is [', CI_lb0, ',', CI_ub0, '].'))
```

```
## [1] "The confidence interval for the intercept is [ 1.44985442731261 , 4.36666156855942 ]."
```

Statistians are interested if 0 is included in the interval for β_1 since they can design hypothesis test to examine whether the response varibale is associated with the explanatory varibale.

d) The hypothesis test is designed as follows

$$H_0 : \beta_1 = 0 \quad H_a : \beta_1 \neq 0$$

```r
Ts <- abs((b1 - 0) / sb1)
pvalue <- (1 - pt(Ts,n-2)) * 2
print(paste('P-vlaue is', pvalue))
```

```
## [1] "P-vlaue is 0.675440753725872"
```

Since p-value is greater than $\alpha = 0.05$, we fail to reject the null hypothesis.

3. When conducting statistical tests concerning the parameter, why is the T test more versatile than the F test?

Solution:

There are main two reseasons: (1) T test can be used to do one-sided test and two-sided test, but F test can only be applied to do two-sided test; (2) T test can be used to test the null hypothesis $H_0 : \beta_1 = \beta_1^*$, where β_1^* can be nonzero values. However, F test can only test null hypothesis $H_0 : \beta_1 = 0$.

4. For each of the following questions, explain whether a confidence interval for a mean response or a prediction interval for a new observation is appropriate.

a) What will be the humidity level in this greenhouse tomorrow when we set the temperature level at 31°C?
b) How much do families whose disposable income is $23,500$ spend, on the average,for meals away from home?
c) How many kilowatt-hours of electricity will be consumed next month by commercial and industrial users in the Twin Cities service area, given that the index of business activity for area remains at its present level?

Solution:

a) A prediction interval for a new observation is required.
b) Confidence interval for a mean response is required.
c) A prediction interval for a new observation is required.

5. A substance used in biological and medical research is shipped by air fright to users in cartons of $1,000$ ampules. The data below, involving 10 shipments, were collected on the number of times the carton was transferred from one aircraft to another over shipment route (X) and the number of ampules found to be broken upon arrival (Y). Assume a simple linear regression is appropriate. Data: Airfreight.csv

a) Verify that the fitted regression line goes through the point $(\bar{X}, \bar{Y})$.
b) Because of changes in airline routes, shipments may have to be transferred more frequently than in the past. Estimate the mean breakage for the following numbers of transfers: $X = 2$. Use a separate 99 percent confidence intervals. Interpret your results.
c) Next shipment will entail two transfers. Obtain a 99 percent prediction interval for the number of broken ampules for this shipment. Interpret your prediction interval.
d) In the next several days, three independent shipments will be made, each entailing two transfers. Obtain a 99 percent prediction interval for the mean number of ampules broken in three shipments.

Solution:

a) The regression line is obtained by R

```r
airfreight_df <- read.csv('data/airfreight.csv', header = TRUE)
names(airfreight_df)
```

```
## [1] "X" "Y"
```

```r
X <- airfreight_df$X
Y <- airfreight_df$Y
SSXY <- t(X - mean(X)) %*% (Y - mean(Y))
SSX <- t(X - mean(X)) %*% (X - mean(X))
```

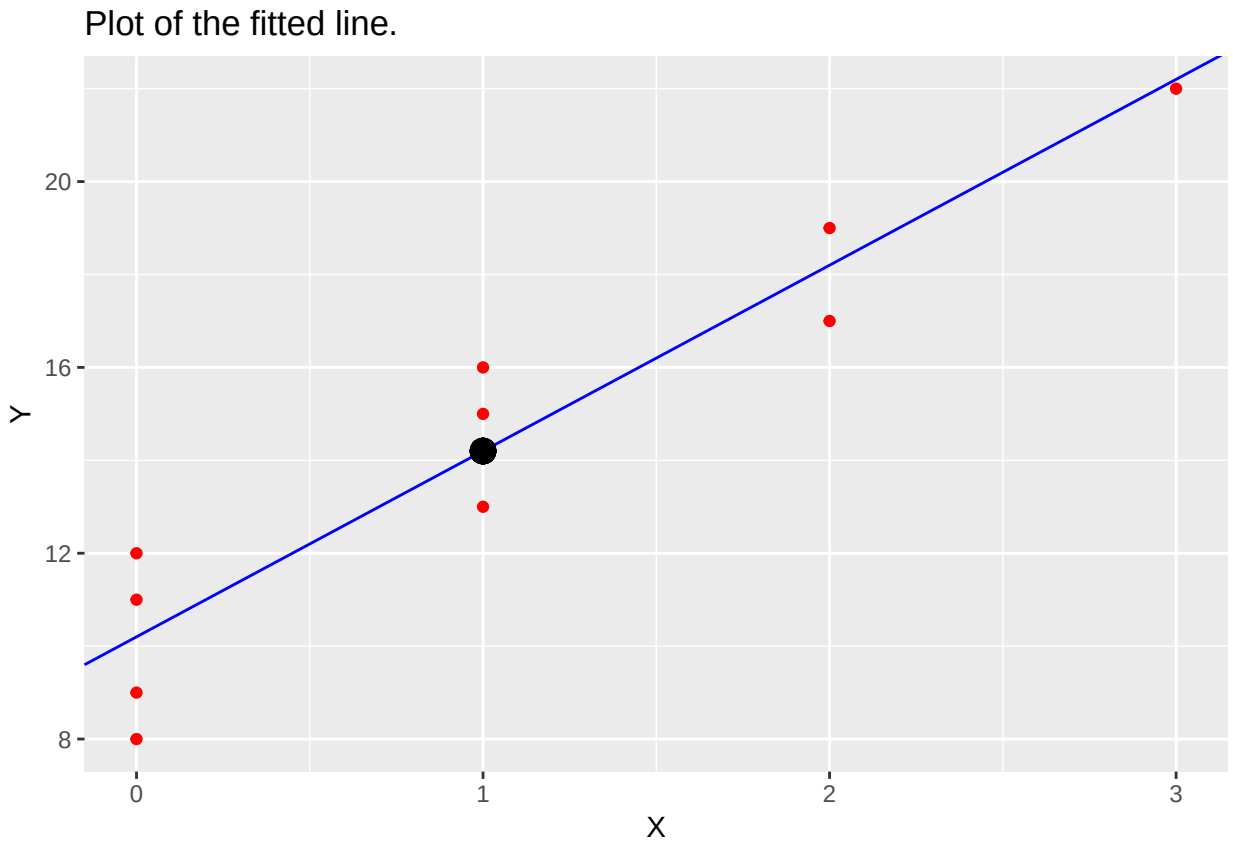
```r
slope <- SSXY / SSX
intercept <- mean(Y) - slope * mean(X)
print(paste('The fitted line is Y =', slope, 'X +', intercept, '.'))
```

```
## [1] "The fitted line is Y = 4 X + 10.2 ."
```

```r
print(paste('The mean value is (', mean(X),',', mean(Y), ').'))
```

```
## [1] "The mean value is ( 1 , 14.2 )."
```

```r
library(ggplot2)
ggplot(data = data.frame(x=X,y=Y), aes(X, Y)) + geom_abline(slope = slope, intercept = intercept, color=
  geom_point(color='red') +
  geom_point(aes(mean(X),mean(Y)), shape=21, fill='black', size=4) +
  ggtitle(paste0('Plot of the fitted line.'))
```



b) The mean breakage for $X = 2$ is $\hat{Y}_i = 18.2$.

```r
Y_hat <- slope * X + intercept
Y_residual <- Y - Y_hat
SSE <- sum(Y_residual^2)
N <- length(X)
MSE <- SSE / (N - 2)
std_Y_hat <- sqrt(MSE * (1 / N + (2 - mean(X))^2 / SSX))
CI_lbYhat <- 18.2 - qt(1 - 0.01/2,N - 2) * std_Y_hat
CI_ubhat <- 18.2 + qt(1 - 0.01/2,N - 2) * std_Y_hat
print(paste('The confidence interval for mean breakage is [', CI_lbYhat, ',', CI_ubhat, '].'))
```

```
## [1] "The confidence interval for mean breakage is [ 15.974287839131 , 20.425712160869 ]."
```

It is concluded that we are 99% certain that this interval contains the mean breakage on $X = 2$.

c) Let us use $\hat{Y}_i = 18.2$ as a point estimate for the predicted number of broken ampules Y_{hnew}.

```
std_Y_hnew <- sqrt(MSE * (1 / N + (2 - mean(X))^2 / SSX + 1))
CI_lbYhnew <- 18.2 - qt(1 - 0.01 / 2, N - 2) * std_Y_hnew
CI_ubYhnew <- 18.2 + qt(1 - 0.01 / 2, N - 2) * std_Y_hnew
cat(paste('The prediction interval for the number of broken ampules for this shipment is\n [', CI_lbYhn
```

```
## The prediction interval for the number of broken ampules for this shipment is
##  [ 12.7481408915637 , 23.6518591084363 ].
```

It is concluded that we are 99% certain that the predicted new number of boken ampules on $X = 2$ will be contained in this interval.

d) To solve for the prediction interval for the mean of three new independent shipments, the estimated standard error for the mean is given by

$$s_{\bar{Y}_{hnew}} = \sqrt{MSE \cdot (\frac{1}{N} + \frac{1}{3} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2})}$$

```
std_Ybar_hnew <- sqrt(MSE * (1 / 3 + 1 / N + (2 - mean(X))^2 / SSX))
CI_lbYbarhnew <- 18.2 - qt(1 - 0.01 / 2, N - 2) * std_Ybar_hnew
CI_ubYbarhnew <- 18.2 + qt(1 - 0.01 / 2, N - 2) * std_Ybar_hnew
cat(paste('The prediction interval for the number of broken ampules for this shipment is\n [', CI_lbYba
```

```
## The prediction interval for the number of broken ampules for this shipment is
##  [ 14.5654272610425 , 21.8345727389575 ].
```

6. Write the name of the following notations, related formulas (if there is one) and describe what they measure.

a) Y
b) $\bar{Y}$
c) $\hat{Y}$
d) ϵ_i (epsilon)
e) e_i
f) β_1
g) b_1
h) σ(residual)
i) s(residual)
j) MSE
k) $s\{b_1\}$
l) $s\{\hat{Y}_h\}$
m) $s\{pred\}$
n) $s\{predmean\}$

Solution:

a) Y is the response variable and it is a random varible.
b) $\bar{Y}$ is the arithmetic mean (average) of the sample response variables.
c) $\hat{Y}$ is the estimated mean of Y, when the predictor is given.
d) ϵ_i is the random error for the $i-th$ observation, we have $\epsilon_i \sim N(0, \sigma^2)$.
e) e_i is the residual for the $i-th$ observation, we have $e_i = Y_i - \hat{Y}_i$.
f) β_1 is the true slope for the simple linear regression model.
g) b_1 is the estimated slope for the simple linear regression model, obtained with least square method.
h) σ(residual) is the standard deviation for the residual.
i) s(residual) is the estimated standard deviation for the residual.

j) MSE is the mean squared error (residual), it is given by

$$MSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)}{n-2}$$

k) $s\{b_1\}$ is the estimated standard deviation for b_1, which is given by

$$s\{b_1\} = \sqrt{\frac{MSE}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

l) $s\{\hat{Y}_h\}$ is the estimated standard deviation for the predicted estimated mean response variable. It is computed by

$$s\{\hat{Y}_h\} = \sqrt{MSE \cdot [\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}]}$$

m) $s\{pred\}$ is the estimated standard error for the new predicted response varibale, it is computed by

$$s\{pred\} = \sqrt{MSE \cdot [1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}]}$$

n) $s\{predmean\}$ is the estimated standard deviation for the mean of m independent new predicted response variables at the same predictors. It is computed by

$$s\{predmean\} = \sqrt{MSE \cdot [\frac{1}{m} + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}]}$$