STA 108 HW3 - J. Jiang

Dylan M Ang

2/1/2022

HW 3: 2.22, 2.25, 2.29, 2.42, 2.51.

2.22

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}, \quad 0 \leq R^2 \leq 1 \tag{1}$$

R^2 is a measure of the fit, when it equals zero, that indicates a weak correlation between terms.

Yes, it is possible that R^2 for the entire data set would be greater than 0, even if the first ten results are zero. In order for this to be the case though, the other 20 experiments would need to have high enough values for R^2 to bring the total over 0.

It wouldn't be possible for the first ten values of R^2 to be greater than 0, but the total be 0. There are already values greater than 0, indicating a relation, so there isn't any way to go back to 0.

2.25 Airplane Breakage

a

```r
break_data = read.table("../datasets/airfreight+breakage.txt")
Y = break_data$V1
X = break_data$V2
n = length(X)
fit = lm(Y ~ X)
kable(anova(fit))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X	1	160.0	160.0	72.72727	2.75e-05
Residuals	8	17.6	2.2	NA	NA

The degrees of freedom and Sum of Squares columns are additive, since $SSTO = SSE + SSR$, and $n - 2 + 1 = n - 1$.

b

Null Hypothesis H_0: $\beta_1 = 0$

Alternate Hypothesis H_1: $\beta_1 \neq 0$

The distribution of F under the null hypothesis is $F_{1,n-2}$. So we reject H_0 when $F > F(1 - \alpha, 1, n - 2)$

```r
Y_hat = fit$fitted.values
SSTO = sum( (Y - mean(Y) )^2 )
SSE = sum((Y - Y_hat)^2)
SSR = sum((Y_hat - mean(Y))^2)
MSR = SSR/1 # df = 1
MSE = SSE/(n - 2)
F_stat = MSR/MSE
pval =pf(F_stat, 1, n - 2, lower.tail = F)
res = data.frame(Source = c("Regression", "Error", "Total"),
                 df = c(1, n-2, n-1),
                 SS = c(SSR, SSE, SSTO),
                 MS = c(MSR, MSE, NA),
                 F_val = c(F_stat, NA, NA),
                 p_val = c(pval, NA, NA)
                 )
kable(res)
```

Source	df	SS	MS	F_val	p_val
Regression	1	160.0	160.0	72.72727	2.75e-05
Error	8	17.6	2.2	NA	NA
Total	9	177.6	NA	NA	NA

The $p < \alpha$ for $\alpha = 0.05$, therefore we reject the null hypothesis. Therefore, We can say with 95% confidence that there is a linear association between the number of carton transfers and the number of broken ampules.

c

$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - 1}{SE(\hat{\beta}_1)} \tag{2}$$

```
## [1] "t = 8.528029"
```

So, $t^2 = 8.528^2 = 72.72727 = F$

d

$R^2 = \frac{SSR}{SSTO}$

```
## [1] "R^2 = 0.900901"
```

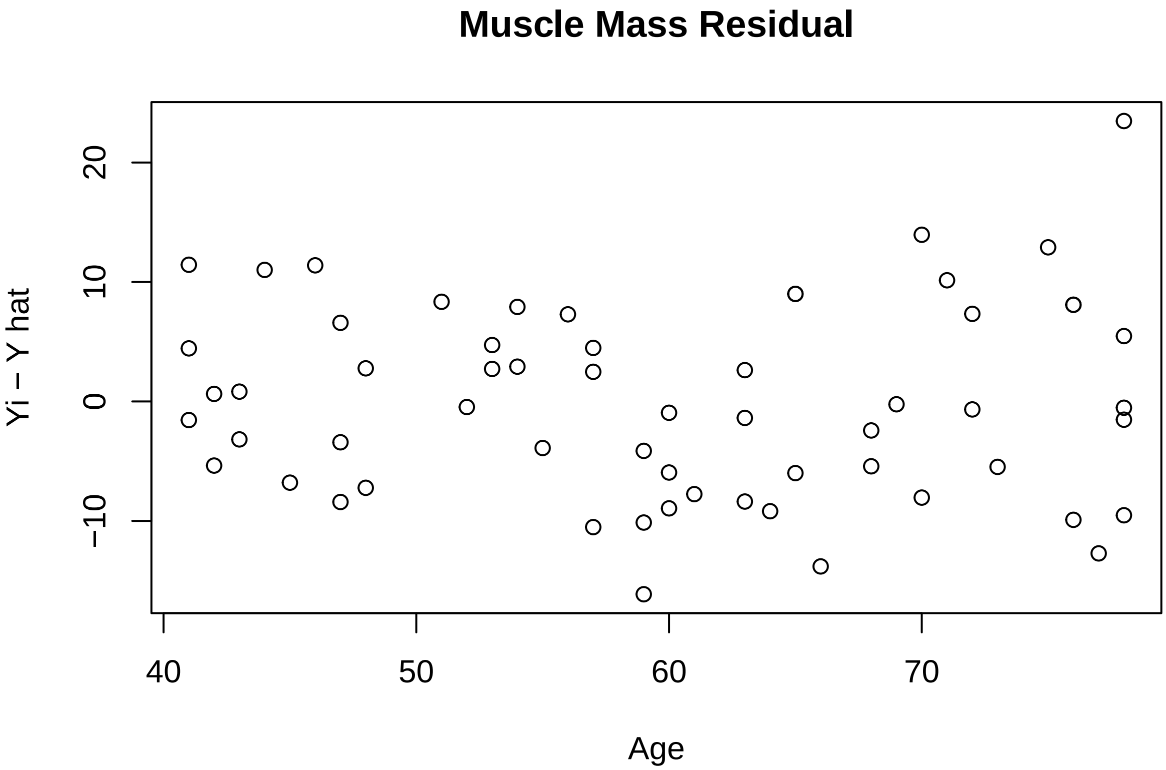
```
## [1] "r = 0.949158"
```

Interpretation: 90% of the variation in Y is accounted for by X.

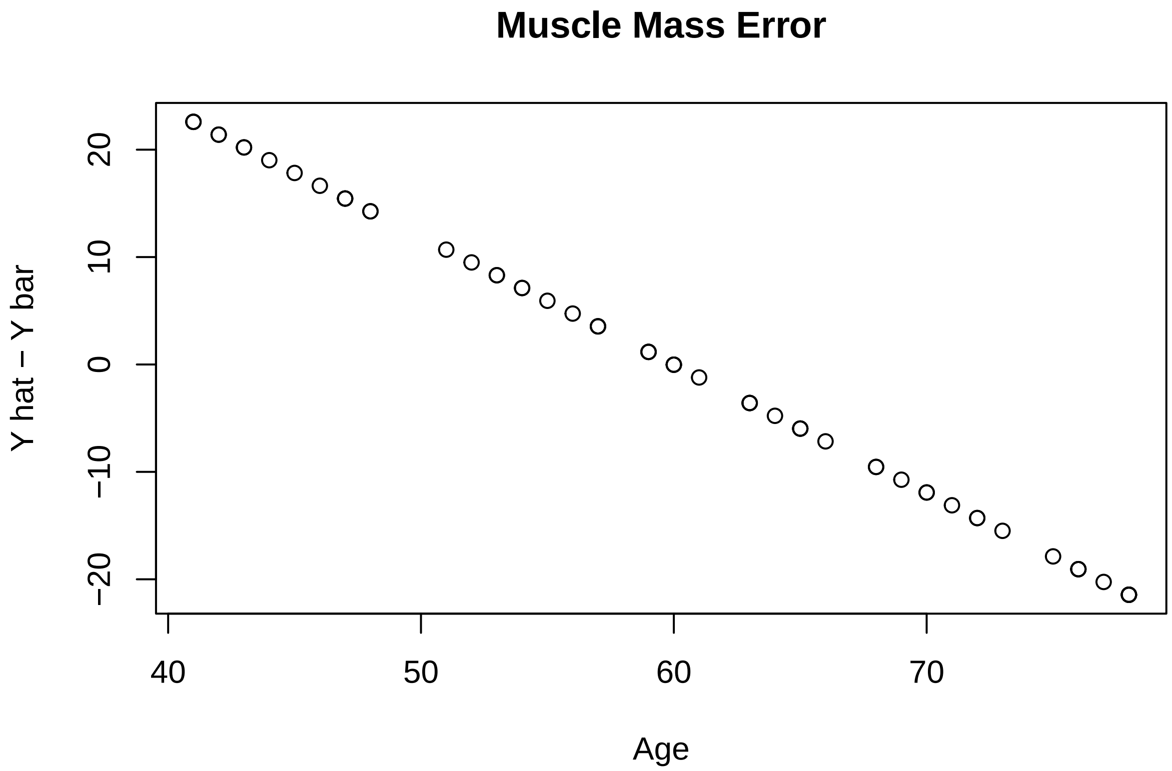
2.29

a Muscle Mass

```
mass = read.table('../datasets/muscle+mass.txt')
Y = mass$V1
X = mass$V2

mass.lm = lm(Y ~ X)

b0hat = mass.lm$coefficients[1]
b1hat = mass.lm$coefficients[2]

yhat = b0hat + b1hat * X

mass.res= Y - yhat
plot(X, mass.res,
     xlab = "Age", ylab = "Yi - Y hat",
     main = "Muscle Mass Residual",
    )
```



```
mass.err = yhat - mean(Y)
plot(X, mass.err,
     xlab = "Age", ylab = "Y hat - Y bar",
     main = "Muscle Mass Error",
     )
```



From the graphs, SSE contributes less to SSTO, and SSR is a larger component. This implies that R^2 is closer to 1, since SSR is a larger portion of SSTO, and $R^2 = \frac{SSR}{SSTO}$.

b

```
kable(anova(mass.lm))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X	1	11627.486	11627.48584	174.062	0
Residuals	58	3874.447	66.80082	NA	NA

c

Null Hypothesis $H_0 : \beta_1 = 0$

Alternate Hypothesis $H_1 : \beta_1 \neq 0$

The distribution of F under the null hypothesis is $F_{1,n-2}$. So we reject H_0 when $F > F(1-\alpha, 1, n-2)$

```
Y_hat = mass.lm$fitted.values
SSTO = sum( (Y - mean(Y) )^2 )
SSE = sum((Y - yhat)^2)
SSR = sum((Y_hat - mean(Y))^2)
MSR = SSR/1 # df = 1
MSE = SSE/(n - 2)
F_stat = MSR/MSE
```

```
pval =pf(F_stat, 1, n - 2, lower.tail = F)
res = data.frame(Source = c("Regression", "Error", "Total"),
                 df = c(1, n-2, n-1),
                 SS = c(SSR, SSE, SSTO),
                 MS = c(MSR, MSE, NA),
                 F_val = c(F_stat, NA, NA),
                 p_val = c(pval, NA, NA)
                 )
kable(res)
```

Source	df	SS	MS	F_val	p_val
Regression	1	11627.486	11627.4858	24.00856	0.001194
Error	8	3874.447	484.3059	NA	NA
Total	9	15501.933	NA	NA	NA

$p < \alpha$ for $\alpha = 0.05$, therefore we can reject the H_0. Therefore, we can say that $\beta_1 \neq 0$.

d

75% of the variation in Y is explained by X, therefore 25% of the variation in muscle mass is not accounted for by age. 75% is a high level of correlation.

e

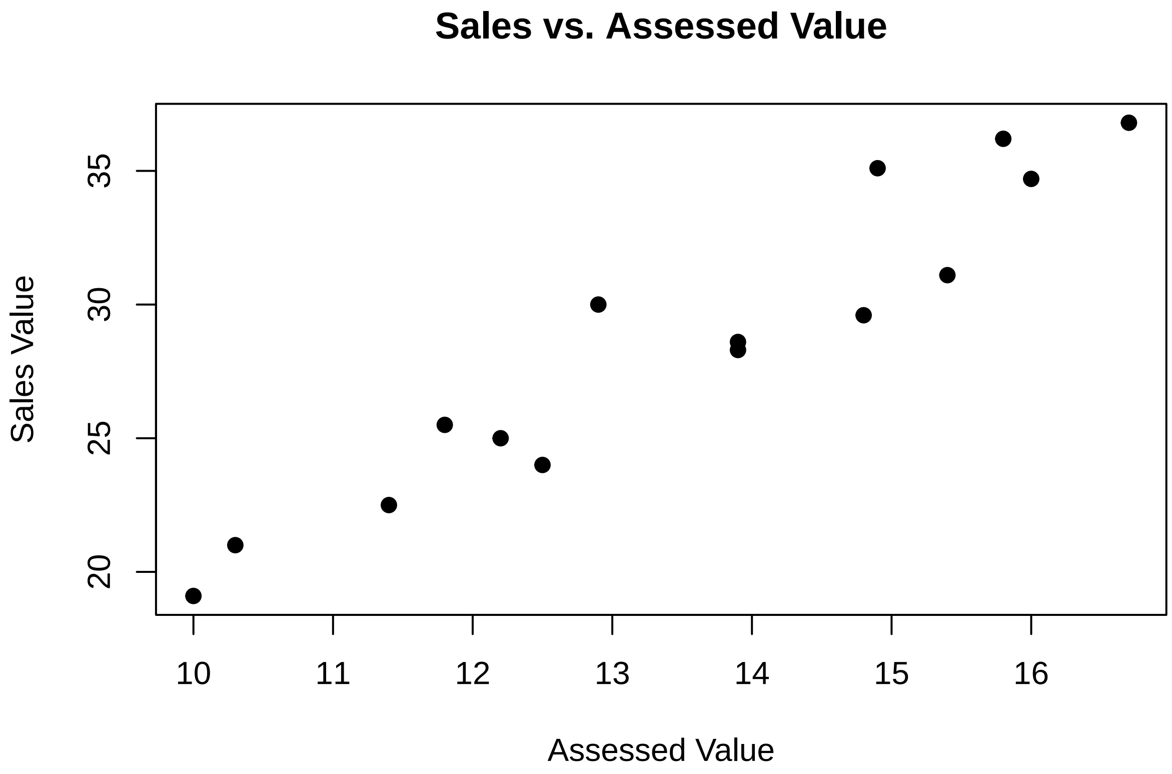
```
## [1] "R^2 = 0.750067"
```

```
## [1] "r = -0.866064"
```

2.42 Property Assessments

a

```
property = read.table('../datasets/property+assessments.txt')
Y1 = property$V1
Y2 = property$V2
n = length(Y1)
plot(Y1, Y2,
     xlab = "Assessed Value",
     ylab = "Sales Value",
     main = "Sales vs. Assessed Value",
     pch=19
     )
```



The two values are closely correlated, therefore a bivariate normal distribution appears to be appropriate.

b

```
r12 = cor(Y1, Y2)
r12
```

```
## [1] 0.9528469
```

r_{12} is the Pearson correlation coefficient, it estimates ρ, the population correlation coefficient.

c

Null Hypothesis $H_0 : \beta_{12} = 0$

Alternate Hypothesis $H_1 : \beta_{12} \neq 0$

$$t^* = \frac{r_{12}\sqrt{n-2}}{\sqrt{1-r_{12}^2}} \tag{3}$$

If $|t^*| \leq t(1 - \alpha/2 \; n-2)$, conclude H_0

If $|t^*| > t(1 - \alpha/2 \; n-2)$, conclude H_1

```
tstar = (r12 * sqrt(n - 2))/(sqrt(1 - r12^2))
crit = qt(p = 0.01/2, df = n - 2, lower.tail=FALSE)
tstar > crit
```

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

$t^* >$ our critical t-value, so we can reject H_0. Therefore, we can say that $\beta_{12} \neq 0$, and $\beta_{21} \neq 0$ since they are equivalent.

d

No, it is not appropriate to use equation 2.87 for ρ.

2.51

Show that $E[b_0] = \beta_0$.

As defined in 2.21,

$$
\begin{aligned}
b_0 &= \bar{Y} - b_1\bar{X} && (4)\\
E[b_0] &= E[\bar{Y} - b_1\bar{X}] && (5)\\
&= E[\bar{Y}] - E[b_1\bar{X}] && (6)\\
&= E[\frac{1}{n}\Sigma_{i=1}^{n}Y_i] - \bar{X}E[b_1] && (7)\\
&= \frac{1}{n}\Sigma_{i=1}^{n}E[Y_i] - \bar{X}\beta_1 && (8)\\
&= \frac{1}{n}\Sigma_{i=1}^{n}E[\beta_0 + \beta_1 x_i] - \bar{X}\beta_1, \quad E(\epsilon_i) = 0 && (9)\\
&= \frac{1}{n}(\Sigma_{i=1}^{n}\beta_0 + \Sigma_{i=1}^{n}\beta_1 x_i] - \bar{X}\beta_1 && (10)\\
&= \frac{1}{n}(n\beta_0 + \beta_1\Sigma_{i=1}^{n}x_i) - \bar{X}\beta_1 && (11)\\
&= \beta_0 + \beta_1\frac{\Sigma_{i=1}^{n}x_i}{n} - \bar{X}\beta_1 && (12)\\
&= \beta_0, \quad \text{definition of } \bar{X} && (13)\\
E[b_0] &= \beta_0 \quad \square && (14)
\end{aligned}
$$