GLIB DOLOTOV (U14734371) MA214 SPRG. 2019

FCAT 11.66 **FCAT scores and poverty.** Refer to the *Journal of Educational and Behavioral Statistics* (Spring 2004) study of scores on the Florida Comprehensive Assessment Test (FCAT), first presented in Exercise 11.30 (p. 600). Consider the simple linear regression relating math score (y) to percentage (x) of students below the poverty level.

1. Test whether y is negatively related to x. Use α=.01.
2. Construct a 99% confidence interval for β1. Interpret the result practically.

1. $H_0: \beta_1 = 0$ Test statistic: $t_c = \frac{\hat\beta_1}{s/\sqrt{SS_{xx}}}$ $t_{0.01,\ df = n-2 =}$

$H_0: \beta_1 < 0$

Estimation of σ^2 for a (First-Order) Straight-Line Model

$$s^2 = \frac{\text{SSE}}{\text{Degrees of freedom for error}} = \frac{\text{SSE}}{n-2}$$

where $\text{SSE} = \sum (y_i - \hat y_i)^2 = \text{SS}_{yy} - \hat\beta_1 \text{SS}_{xy}$

in which

$$\text{SS}_{yy} = \sum (y_i - \bar y)^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

To estimate the standard deviation σ of ε, we calculate

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

We will refer to s as the **estimated standard error of the regression model.**

Test statistic: $$t_c = \frac{\hat\beta_1}{s_{\hat\beta_1}} = \frac{\hat\beta_1}{(s/\sqrt{\text{SS}_{xx}})}$$

$$s^2 = \frac{SSE}{n-2} = \frac{SS_{yy} - SS_{xy} \cdot \hat\beta_1}{n-2}$$

```
> data <- read.csv("FCATtxt.csv")
> # Using R
> model <- lm(data$MATH ~ data$POVERTY)
> summary(model)

Call:
lm(formula = data$MATH ~ data$POVERTY)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9020  -2.4388   0.3001   2.7826  11.1925

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  189.81582    3.02148  62.822  < 2e-16 ***
data$POVERTY  -0.30544    0.04759  -6.418 2.93e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.366 on 20 degrees of freedom
Multiple R-squared:  0.6731,	Adjusted R-squared:  0.6568
F-statistic: 41.19 on 1 and 20 DF,  p-value: 2.927e-06
```

```
> # Manually
> attach(data)
The following objects are masked from data (pos = 3):

    MATH, POVERTY, READING

> b_1 <- sum( (MATH - mean(MATH)) * (POVERTY - mean(POVERTY)) ) /
+   sum( (POVERTY - mean(POVERTY))^2 )
> b_1
[1] -0.3054445
> s2 <-
+   (
+     sum( (MATH - mean(MATH))^2 ) -
+     b_1 * sum( (MATH - mean(MATH)) * (POVERTY - mean(POVERTY)) )
+   ) / (length(MATH) - 2)
> s <- sqrt(s2)
> s
[1] 5.365722
> s_b_1 <- s / sqrt( sum((POVERTY - mean(POVERTY))^2) )
> s_b_1
[1] 0.04759332
> t_stat <- b_1 / s_b_1
> t_stat
[1] -6.417802
> qt(0.99, length(MATH) - 2)
[1] 2.527977
```

$\hat\beta_1 = \frac{SS_{xy}}{SS_{xx}} = -0.3054$

$s^2 = \frac{SSE}{n-2} = \frac{SS_{yy} - \hat\beta_1 SS_{xy}}{n-2}$

$s = \sqrt{s^2} = 5.366$

$s_{\hat\beta_1} = \frac{s}{\sqrt{SS_{xx}}} = 0.0476$

$t_c = \frac{\hat\beta_1}{s_{\hat\beta_1}} = \frac{-0.3054}{0.0476} = -6.418$

t_α

$t_c = -6.418 < -t_\alpha = -2.528$ $\Rightarrow$ reject null in favor of alternate hyp, $\beta_1 < 0$.

2) $\hat\beta_1 \pm (t_{\alpha/2}) s_{\hat\beta_1}$ $= -0.3054 \pm t_{0.005} \cdot 0.04759$

$= -0.3054 \pm 2.8453 \cdot 0.04759$

```
> qt(0.995, length(MATH) - 2)
[1] 2.84534
```

$= \boxed{-0.3054 \pm 0.1354}$

W/ 99% level of confidence, β_1 is located w/i the confidence interval. Note that the entire interval is negative. $\Rightarrow \beta_1$ is negative

OJUICE 11.50 **Sweetness of orange juice.** Refer to the study of the quality of orange juice produced at a juice manufacturing plant, Exercise 11.32 (p. 601). Recall that simple linear regression was used to predict the sweetness index (y) from the amount of pectin (x) in orange juice manufactured during a production run.

1. Give the values of SSE, $s2$, and s for this regression.
2. Explain why it is difficult to give a practical interpretation to $s2$.
3. Use the value of s to derive a range within which most (about 95%) of the errors of prediction of sweetness index fall.

1. $$SSE = SS_{yy} - \hat{\beta}_1 SS_{xy}$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \Rightarrow SSE = SS_{yy} - (SS_{xy})^2 / SS_{xx}$$

$$SSE = 1.3183 - (-130.442)^2/(56452.96) = 1.0169$$

$$s^2 = \frac{SSE}{n-2} = \frac{1.0169}{24-2} = 0.0462$$

$$s = \sqrt{s^2} = \sqrt{0.0462} = 0.2150.$$

```
> # 2)
> data <- read.csv("OJUICEtxt.csv")
> attach(data)
> summary(lm(SweetIndex ~ Pectin))

Call:
lm(formula = SweetIndex ~ Pectin)

Residuals:
     Min       1Q   Median       3Q      Max
-0.54373 -0.11039  0.06089  0.13432  0.34638

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.2520679  0.2366220  26.422   <2e-16 ***
Pectin      -0.0023106  0.0009049  -2.554   0.0181 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.215 on 22 degrees of freedom
Multiple R-squared:  0.2286,    Adjusted R-squared:  0.1936
F-statistic:  6.52 on 1 and 22 DF,  p-value: 0.01811
```

```
> n <- length(SweetIndex)
> n
[1] 24
> ssxx <- sum( (Pectin - mean(Pectin))^2 )
> ssxx
[1] 56452.96
> ssyy <- sum( (SweetIndex - mean(SweetIndex))^2 )
> ssyy
[1] 1.318333
> ssxy <- sum( (Pectin - mean(Pectin)) * (SweetIndex - mean(SweetIndex)))
> ssxy
[1] -130.4417
```

$$\boxed{SSE = 1.0169 \quad s^2 = 0.0462 \quad s = 0.2150}$$

2. s^2 is hard to interperet because of its units, which are usually some measurement squared.

3. **Interpretation of s, the Estimated Standard Deviation of ε**

We expect most ($\approx 95\%$) of the observed y values to lie within $2s$ of their respective least squares predicted values, $\hat{y}$.

95% of values $\hat{y} - y_i$ will fall w/i the range $(-2s, 2s)$

$$= (-2 \cdot 0.2150,\ 2 \cdot 0.2150) = \boxed{(-0.430,\ 0.430)}$$

BBALL 11.116 **Sound waves from a basketball.** Refer to the *American Journal of Physics* (June 2010) study of sound waves in a spherical cavity, Exercise 11.31 (p. 601). You fit a straight-line model relating sound wave frequency (y) to number of resonances (x) for

$n=24$

echoes resulting from striking a basketball with a metal rod.

1. Use the model to predict the sound wave frequency for the 10th resonance.
2. Form a 90% confidence interval for the prediction, part **a.** Interpret the result.
3. Suppose you want to predict the sound wave frequency for the 30th resonance. What are the dangers in making this prediction with the fitted model?

```
> data <- read.csv("BBALLtxt.csv")
> attach(data)
The following objects are masked from data (pos = 3):

    Frequency, Resonance

> x_bar <- mean(Resonance)
> y_bar <- mean(Frequency)
> ssxx <- sum( (Resonance - x_bar)^2 )
> ssyy <- sum( (Frequency - y_bar)^2 )
> ssxy <- sum( (Resonance - x_bar) * (Frequency - y_bar))
> n <- length(Resonance)
> setNames(c(x_bar, y_bar, n, ssxx, ssyy, ssxy),
+          c("x_bar", "y_bar", "n","ssxx", "ssyy", "ssxy"))
      x_bar        y_bar            n         ssxx         ssyy         ssxy
     12.500     4103.917       24.000     1150.000 52354781.833   242380.000
```

1. $\hat{\beta}_1 = \dfrac{SS_{xy}}{SS_{xx}} = \dfrac{242380}{1150}$

$\hat{\beta}_1 = 210.765$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 4103.917 - 210.765 \cdot 12.5$

$\hat{\beta}_0 = 1469.352$

$\hat{y}_{10} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_{10}$

$= 1469.352 + 210.765 \cdot 10$

$\boxed{\hat{y}_{10} = 3577.004}$

```
> summary(lm(Frequency ~ Resonance))

Call:
lm(formula = Frequency ~ Resonance)

Residuals:
    Min      1Q  Median      3Q     Max
-701.12 -134.49   69.35  164.67  275.53

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1469.351    101.216   14.52 9.47e-13 ***
Resonance    210.765      7.084   29.75  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 240.2 on 22 degrees of freedom
Multiple R-squared:  0.9758,	Adjusted R-squared:  0.9746
F-statistic: 885.3 on 1 and 22 DF,  p-value: < 2.2e-16
```

2. $\hat{y} \pm t_{0.05,\ df=22} \cdot s \cdot \sqrt{\dfrac{1}{n} + \dfrac{(x_p - \bar{x})^2}{SS_{xx}}}$

$s = \sqrt{\dfrac{SS_{yy} - \hat{\beta}_1 \cdot SS_{xy}}{n-2}}$

$= \sqrt{\dfrac{52354781.8 - 210.765 \cdot 242380}{22}}$

$s = 240.2185$

```
> b_1 <- ssxy / ssxx
> s <- sqrt( (ssyy - b_1 * ssxy) / (n - 2))
> s
[1] 240.2185
```

$t_{0.05,\ df=22} = 1.717$

```
> qt(0.95, df = 22)
[1] 1.717144
```

A 100(1 − α)% Confidence Interval for the Mean Value of y at x = x_p

$$\hat{y} \pm t_{\alpha/2}(\text{Estimated standard error of } \hat{y})$$

or

$$\hat{y} \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

where $t_{\alpha/2}$ is based on $(n-2)$ degrees of freedom.

```
> predict(model, newdata = data.frame(Resonance=10), interval = "confidence", level = 0.90)
       fit      lwr      upr
1 3577.004 3487.481 3666.526
```

A 100(1 − α)% Prediction Interval* for an Individual New Value of y at x = x_p

$$\hat{y} \pm t_{\alpha/2}(\text{Estimated standard error of prediction})$$

or

$$\hat{y} \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

where $t_{\alpha/2}$ is based on $(n-2)$ degrees of freedom.

$CI = 1680.117 \pm 1.717 \cdot 240.2185 \cdot \sqrt{\dfrac{1}{24} + \dfrac{(10-12.5)^2}{1150}}$

$\boxed{CI = 3577.004 \pm 89.515}$ → $\approx (3487.5, 3666.5)$

3) The model might not reflect actuality, relationships are rarely linear across their entire span.