

Name:

STAT 516 – Fall 2026
Exam 1

1. A simple linear regression was fit (based on $n = 8$ observations) in order to predict sugar content (Y , in grams) of fruit based on the number of days (X) after the fruit had been picked. Look at the partial R output given below:

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.0419      2.5527   4.717  0.00327 **
days        -0.8216      0.5105  -1.609  0.15867
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.803 on 6 degrees of freedom
Multiple R-squared:  0.3015,    Adjusted R-squared:  0.1851
F-statistic:  2.59 on 1 and 6 DF,  p-value: 0.1587

```

(a) State the estimated regression equation, using the correct notation.

$$\hat{Y} = 12.0419 - 0.8216 X$$

(b) Use the estimated regression equation to predict the sugar content of a fruit that was picked 4 days ago.

$$\hat{Y} = 12.0419 - 0.8216(4) = 8.756 \text{ grams}$$

(c) Find a 90% prediction interval for the sugar content of a fruit that was picked 4 days ago.

Note that $\bar{X} = 4.25$ for this data set. [Hint: Recall that the standard error of $\hat{\beta}_1$ is $\sqrt{\frac{MSE}{S_{xx}}}$, and that what R calls "Residual standard error" is merely $\sqrt{MSE}$.]

$$MSE = 3.803^2 = 14.4628, \text{ and so } 0.5105 = \sqrt{\frac{14.4628}{S_{xx}}}$$

$$\Rightarrow 0.26061 = \frac{14.4628}{S_{xx}} \Rightarrow S_{xx} = \frac{14.4628}{0.26061} \Rightarrow S_{xx} = 55.496$$

$$t_{.05, 6} = 1.943$$

$\Rightarrow 90\% \text{ PI} :$

$$8.756 \pm 1.943 \sqrt{14.4628 \left[1 + \frac{1}{8} + \frac{(4 - 4.25)^2}{55.496} \right]}$$

$$\Rightarrow 8.756 \pm 1.943(4.0357)$$

$$\Rightarrow (0.9146, 16.597)$$

(e) Provide a precise interpretation for the r^2 value given on the output.

$r^2 = 0.3015$. About 30% of the sample variation in sugar content is explained by its linear relationship with number of days.

(f) What is the sample correlation coefficient r between X and Y here? Show your work.

Note $\sqrt{0.3015} = 0.549$, and since the estimated slope is negative, $r = -0.549$

2. A regression analysis was done on 34 apartment complexes to predict the advertised monthly rent (in dollars) based on the complex's age (in years), size (in square feet), security deposit (in dollars), and number of units in the complex. A partial R output is given below:

```
lm(formula = rent ~ age + size + sec_deposit + units)
```

Residuals:

Min	1Q	Median	3Q	Max
-81.479	-28.672	-7.442	16.184	136.896

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	144.36661	120.37630	1.199	0.240121
age	-9.34324	2.89341	-3.229	0.003080 **
size	0.63628	0.16153	3.939	0.000472 ***
sec_deposit	-0.53787	0.33565	-1.602	0.119887
units	0.15078	0.08184	1.842	0.075654 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50.56 on 29 degrees of freedom
Multiple R-squared: 0.4892, Adjusted R-squared: 0.4188
F-statistic: 6.944 on 4 and 29 DF, p-value: 0.0004761

(a) What is $\hat{\beta}_2$, the coefficient of size? Interpret this number carefully in terms of the variables in the problem.

$$\hat{\beta}_2 = 0.6363$$

We estimate that the expected rent increases by 0.6363 dollars for each one sq.ft. increase in size, adjusting for age, security deposit, and units.

(b) Based on the overall F-test, is the set of independent variables significant, as a whole? Please use $\alpha = 0.05$ and provide justification for your answer.

Reject $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ at $\alpha = .05$

$$P\text{-value} = .00048 < .05$$

$$\text{And } F^* = 6.944 > 2.7 \approx F_{.05, 4, 29}$$

Conclude the set of independent variables is not all useless.

(c) Based on the t-test results given, is "age" a significant predictor, given the presence of the other variables in the model? Is "security deposit" a significant predictor, given the presence of the other variables in the model? Be specific and provide some type of numerical justification for your answer (You can use $\alpha = 0.05$).

$$t_{.025, 29} \approx 2.04$$

$|-3.229| > 2.04$, so age is a significant predictor at $\alpha = .05$ (P-value = .003 < .05)

$|-1.602| < 2.04$, so security deposit is not a significant predictor at $\alpha = .05$ (P-value = .1199 > .05)

(d) Another analyst claims that this complicated multiple regression model is not needed and he can predict rent just as well using a linear regression model where "age" and "size" are the only independent variables. The other analyst did this regression and got an SSE of 88775. Conduct an F-test (using $\alpha = 0.05$) to determine if this reduced model is sufficient. State the null hypothesis of the F-test, give the value of the test statistic F^* , and give your conclusion. [Hint: Recall that what R calls "Residual standard error" is merely $\sqrt{MSE}$, and the table value $F_{.05}(2, 29) = 3.33$].

$$H_0: \beta_3 = \beta_4 = 0$$

$$MSE = 50.56^2 = 2556.3$$

$$\Rightarrow SSE = (29)(2556.3) = 74132.7$$

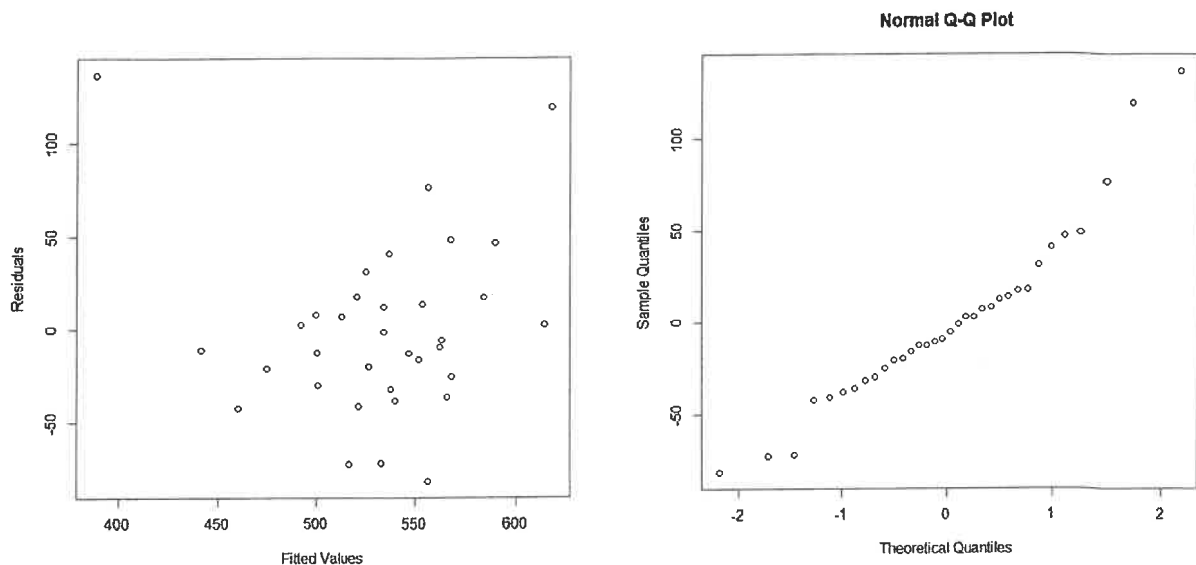
$$F^* = \frac{[88775 - 74132.7] / 2}{2556.3} = 2.86$$

Since $2.86 < 3.33$, we fail to reject H_0 . Conclude that the reduced model with only "age" and "size" as predictors is sufficient.

(e) The following two images give the plot of the residuals against the fitted values and a normal Q-Q plot of the residuals. From these plots, what is the most notable worrisome feature, and what possible violation of the model assumptions does this indicate?

Megaphone shape in residual plot.

- Non-constant error variance



3. In multiple regression, why is multicollinearity a problem? Be as specific as you can.

- Can lead to large variances/std. errors for the estimated coefficients.
- Coefficient estimates not trustworthy / not significant

4. For a data set of pro basketball players, a regression was performed with the dependent variable being annual salary (Y) in millions of dollars and the independent variable being points per game (X). A 90% prediction interval for the Y-value for a new player with $X = 7.3$ was found to be (2.18, 5.56).

(a) What is the best interpretation of this prediction interval?

- (A) With 90% confidence, the mean salary for players who score 7.3 points per game is between 2.18 and 5.56 million dollars.
- (B) About 90% of the pro basketball players in the sample have a salary of between 2.18 and 5.56 million dollars.
- ☒ (C) With probability 0.90, the salary for a new player who scores 7.3 points per game is between 2.18 and 5.56 million dollars.
- (D) With 90% confidence, the change in expected salary for a one-point increase in scoring is between 2.18 and 5.56 million dollars.

(b) If you had to predict *with a single number* the salary of a player who scores 7.3 points per game based on this regression analysis, what would the answer be?

- ☒ (A) \$3.87 million
- (B) \$ 2.18 million
- (C) \$5.56 million
- (D) \$7.3 million

(c) In this data set, what is a possible value for the 90% confidence interval for the mean salary of all players scoring 7.3 points per game? (Choose one)

- (A) (3.91, 4.83)
- (B) (1.37, 6.37)
- ☒ (C) (3.41, 4.33)
- (D) (2.18, 5.56)

5. Choose the best answer: If for two variables X and Y , the correlation coefficient $r > 0$, then as X increases, Y tends to:

- ☒ (A) increase (B) decrease (C) remain constant

4

6. A transformation of the response variable Y may be a remedy for (choose the best answer):

- (A) non-constant error variance
(B) multicollinearity
(C) non-normal errors
(D) both (A) and (B)
☒ (E) both (A) and (C)
(F) all of (A), (B), and (C)

4

Extra credit: Two principles we have studied are: “ correlation ” does not imply causation” and “ extrapolation ” is dangerous when predicting using regression”. How would you fill in the blanks?

+4