Friday, April 19, 2019

Time at the lunch table	Caloric intake
21.4	472
30.8	498
37.7	335
32.8	423
39.5	437
22.8	508
34.1	431
33.9	479
43.8	454
42.4	450
43.1	410
29.2	504
31.3	437
28.6	489
32.9	436
30.6	480
35.1	439
33.0	444
43.7	408

- Warm-up

  $\widehat{calories} = 575.319 - 3.706time$

  - If another student ate in 25 minutes and consumed 400 calories, what is his residual? Interpret it in context.
- More with Linear Regression inference

Objectives

Content Objective: I will use the linear regression analysis to create confidence intervals and perform hypothesis testing.

Social Objective: I will participate in class activities.

Language Objective: I will clearly write down formulas, conditions & assumptions, and other notes.

Warm-up

$$\widehat{calories} = 575.319 - 3.706time$$

If another student ate in 25 minutes and consumed 400 calories, what is his residual? Interpret it in context.

$\widehat{calories} = 575.319 - 3.706(25)$

$= 482.669$

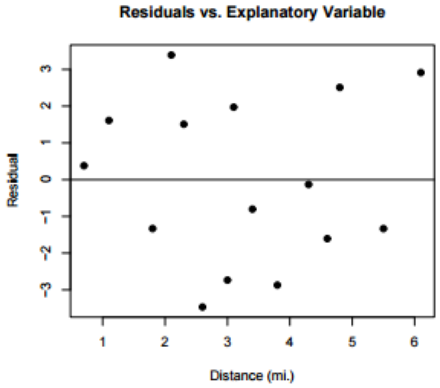
$400 - 482.669 = -82.669$ calories

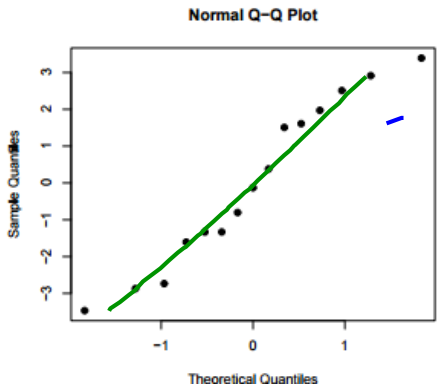
He consumed 82.669 calories less than his time would have predicted using the model.

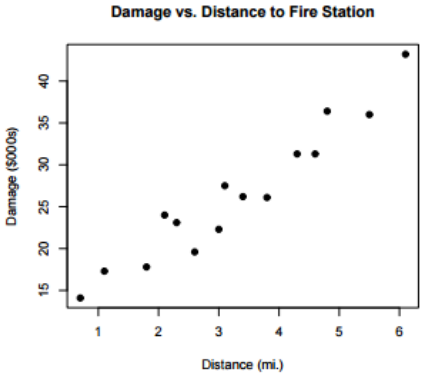
Time at the lunch table	Caloric intake
21.4	472
30.8	498
37.7	335
32.8	423
39.5	437
22.8	508
34.1	431
33.9	479
43.8	454
42.4	450
43.1	410
29.2	504
31.3	437
28.6	489
32.9	436
30.6	480
35.1	439
33.0	444
43.7	408

Suppose a fire insurance company wants to relate the amount of fire damage in major residential fires to the distance between the burning house and the nearest fire station. The study is to be conducted in a large suburb of a major city; a sample of 15 recent fires in this suburb is selected. The amount of damage (in thousands of dollars) and the distance (in miles) between the fire and the nearest fire station are recorded in each fire.

Obs.	Dist.	Damage
1	0.7	14.1
2	1.1	17.3
3	1.8	17.8
4	2.1	24.0
5	2.3	23.1
6	2.6	19.6
7	3.0	22.3
8	3.1	27.5
9	3.4	26.2
10	3.8	26.1
11	4.3	31.3
12	4.6	31.3
13	4.8	36.4
14	5.5	36.0
15	6.1	43.2








```
------------------------------------------------------------------------------
  damage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
distance |   4.919331   .3927478    12.525   0.000     4.070851    5.767811
   _cons |   10.27793   1.420278     7.237   0.000     7.209605    13.34625
------------------------------------------------------------------------------
```

Conditions

Quantitative? yes both miles & money

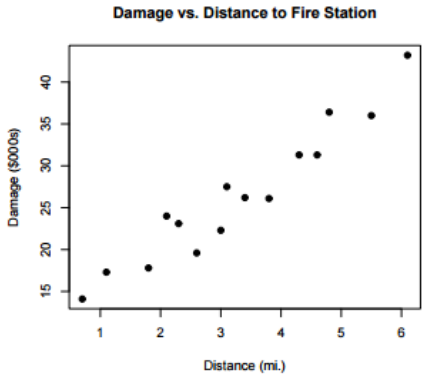
Scatterplot shows a linear pattern (straight enough)

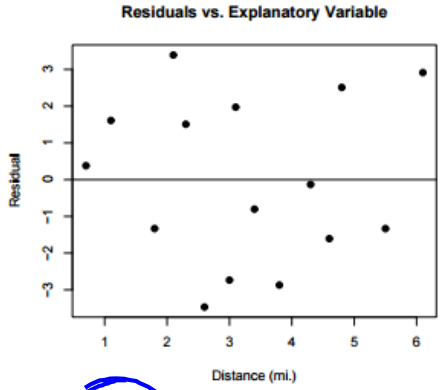
No pattern in the residual plot

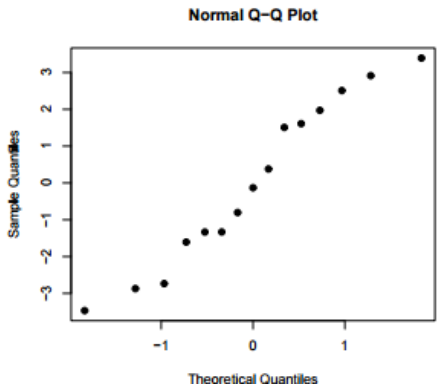
Normal probability plot verifies Nearly Normal

Suppose a fire insurance company wants to relate the amount of fire damage in major residential fires to the distance between the burning house and the nearest fire station. The study is to be conducted in a large suburb of a major city; a sample of 15 recent fires in this suburb is selected. The amount of damage (in thousands of dollars) and the distance (in miles) between the fire and the nearest fire station are recorded in each fire.

Obs.	Dist.	Damage
1	0.7	14.1
2	1.1	17.3
3	1.8	17.8
4	2.1	24.0
5	2.3	23.1
6	2.6	19.6
7	3.0	22.3
8	3.1	27.5
9	3.4	26.2
10	3.8	26.1
11	4.3	31.3
12	4.6	31.3
13	4.8	36.4
14	5.5	36.0
15	6.1	43.2







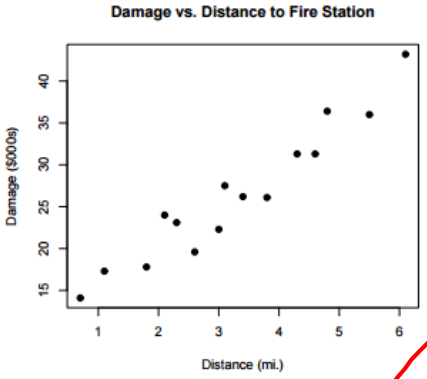
damage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
distance	4.919331	.3927478	12.525	0.000	4.070851	5.767811
_cons	10.27793	1.420278	7.237	0.000	7.209605	13.34625

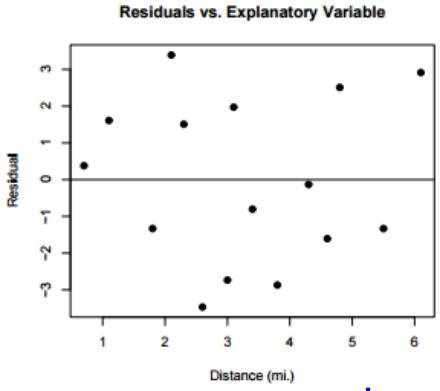
$\widehat{damage} = 10.278 + 4.919\ distance$

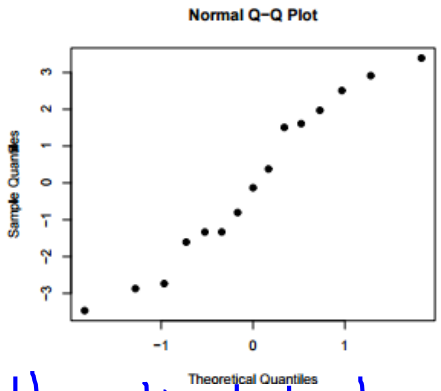
EQUATION OF LINEAR REGRESSION

Suppose a fire insurance company wants to relate the amount of fire damage in major residential fires to the distance between the burning house and the nearest fire station. The study is to be conducted in a large suburb of a major city; a sample of 15 recent fires in this suburb is selected. The amount of damage (in thousands of dollars) and the distance (in miles) between the fire and the nearest fire station are recorded in each fire.

Obs.	Dist.	Damage
1	0.7	14.1
2	1.1	17.3
3	1.8	17.8
4	2.1	24.0
5	2.3	23.1
6	2.6	19.6
7	3.0	22.3
8	3.1	27.5
9	3.4	26.2
10	3.8	26.1
11	4.3	31.3
12	4.6	31.3
13	4.8	36.4
14	5.5	36.0
15	6.1	43.2







Never the adjusted value

damage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
distance	4.919331	.3927478	12.525	0.000	4.070851	5.767811
_cons	10.27793	1.420278	7.237	0.000	7.209605	13.34625

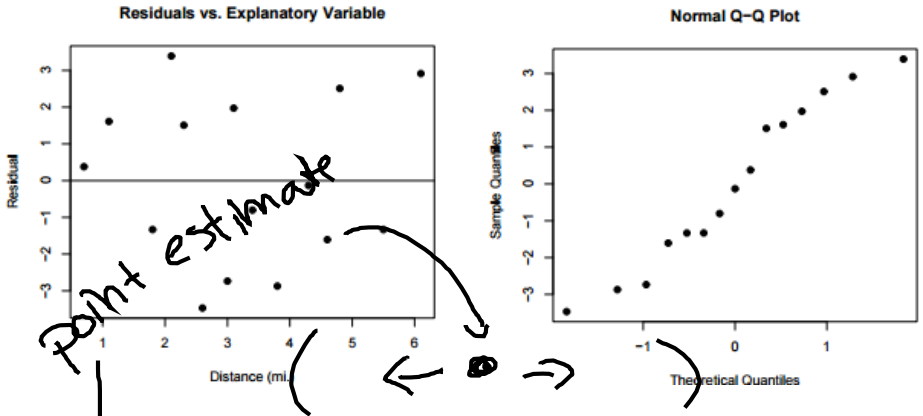
R^2

% of variation in y (damage) that can be predicted by the linear relationship with x (distance).

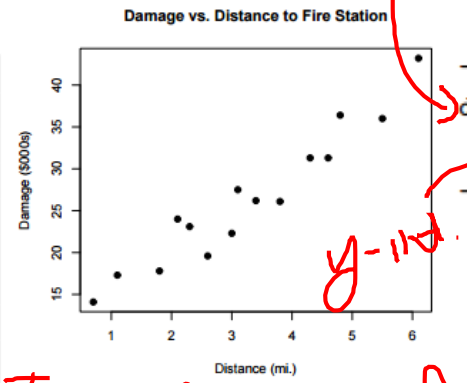
"Coefficient of Determination"

Suppose a fire insurance company wants to relate the amount of fire damage in major residential fires to the distance between the burning house and the nearest fire station. The study is to be conducted in a large suburb of a major city; a sample of 15 recent fires in this suburb is selected. The amount of damage (in thousands of dollars) and the distance (in miles) between the fire and the nearest fire station are recorded in each fire.

Obs.	Dist.	Damage
1	0.7	14.1
2	1.1	17.3
3	1.8	17.8
4	2.1	24.0
5	2.3	23.1
6	2.6	19.6
7	3.0	22.3
8	3.1	27.5
9	3.4	26.2
10	3.8	26.1
11	4.3	31.3
12	4.6	31.3
13	4.8	36.4
14	5.5	36.0
15	6.1	43.2



slope



damage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
distance	4.919331	.3927478	12.525	0.000	4.070851	5.767811
_cons	10.27793	1.420278	7.237	0.000	7.209605	13.34625

Confidence Interval

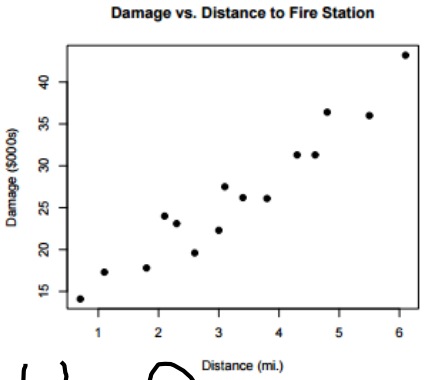
I am 95% confident that the true slope predicting damage from distance is between 4.07 and 5.76.

4.92 - 4.07

ME: 0.85

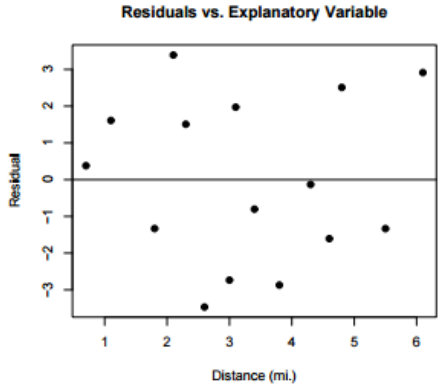
Suppose a fire insurance company wants to relate the amount of fire damage in major residential fires to the distance between the burning house and the nearest fire station. The study is to be conducted in a large suburb of a major city; a sample of 15 recent fires in this suburb is selected. The amount of damage (in thousands of dollars) and the distance (in miles) between the fire and the nearest fire station are recorded in each fire.

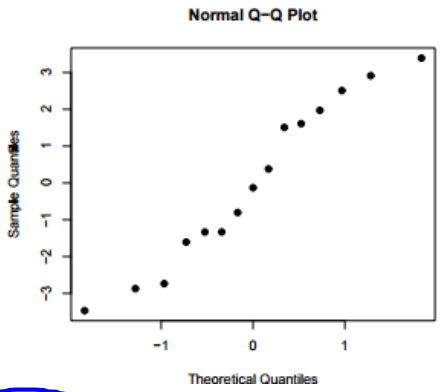
Obs.	Dist.	Damage
1	0.7	14.1
2	1.1	17.3
3	1.8	17.8
4	2.1	24.0
5	2.3	23.1
6	2.6	19.6
7	3.0	22.3
8	3.1	27.5
9	3.4	26.2
10	3.8	26.1
11	4.3	31.3
12	4.6	31.3
13	4.8	36.4
14	5.5	36.0
15	6.1	43.2



$H_0: \beta_1 = 0$

$H_A: \beta_1 > 0$





damage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
distance	4.919331	.3927478	12.525	0.000	4.070851	5.767811
_cons	10.27793	1.420278	7.237	0.000	7.209605	13.34625

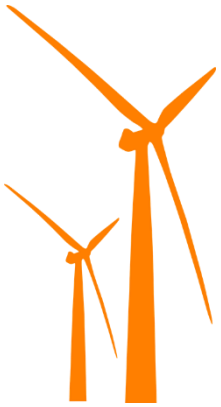
HYPOTHESIS TEST

$t_{13} = 12.525$

p-value ≈ 0

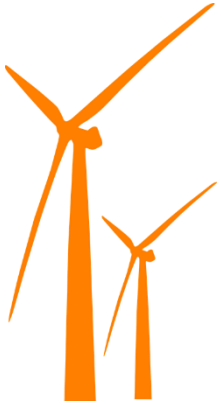
Due to a very low p-value of almost zero, which is $< \alpha$, reject the null. There is sufficient evidence of a positive slope relating distance & damage

Windmills generate electricity by transferring energy from wind to a turbine. A study was conducted to examine the relationship between wind velocity in miles per hour (mph) and electricity production in amperes for one particular windmill. For the windmill, measurements were taken on twenty-five randomly selected days, and the computer output for the regression analysis for predicting electricity production based on wind velocity is given below. The regression model assumptions were checked and determined to be reasonable over the interval of wind speeds represented in the data, which were from 10 miles per hour to 40 miles per hour.



Predictor	Coef	SE Coef	T	P
Constant	0.137	0.126	1.09	0.289
Wind velocity	0.240	0.019	12.63	0.000

S = 0.237 R-Sq = 0.873 R-Sq (adj) = 0.868



(a) Use the computer output above to determine the equation of the least squares regression line. Identify all variables used in the equation.

(b) How much more electricity would the windmill be expected to produce on a day when the wind velocity is 25 mph than on a day when the wind velocity is 15 mph? Show how you arrived at your answer.

(c) What proportion of the variation in electricity production is explained by its linear relationship with wind velocity?

(d) Is there statistically convincing evidence that electricity production by the windmill is related to wind velocity? Explain.

Intent of Question

The primary goals of this question were to assess students' ability to (1) determine the equation of the least squares regression line from a computer output; (2) use the slope of the least squares line to compare expected values of the response variable for different values of the explanatory variable; (3) recognize how to determine the proportion of variability in the response variable explained by the least squares line; (4) use computer output to determine whether the linear relationship between two quantitative variables is statistically significant.

y-int

Predictor	Coef	SE Coef	T	P
Constant	0.137	0.126	1.09	0.289
Wind velocity	0.240	0.019	12.63	0.000

S = 0.237 R-Sq = 0.873 R-Sq (adj) = 0.868

$\widehat{\text{electricity}} = 0.137 + 0.240\,\text{velocity}$

(a) Use the computer output above to determine the equation of the least squares regression line. Identify all variables used in the equation.

The equation of the least squares regression line is

predicted electricity production = 0.137 + 0.240 × wind velocity.

Essentially correct (E) if the response gives the correct equation *AND* includes the following two components:

1. Provides correct variable names (with context).
2. Uses a modifier such as "expected" or "predicted" or "estimated" (or a "hat" symbol) with the response variable, electricity production.

Partially correct (P) if the response gives the correct equation *AND* includes exactly one of the two components listed above.

Incorrect (I) if the response does not meet the criteria for E or P.

(b) How much more electricity would the windmill be expected to produce on a day when the wind velocity is 25 mph than on a day when the wind velocity is 15 mph? Show how you arrived at your answer.

$25 \times 0.240 - 15 \times 0.240$

The slope coefficient of 0.240 indicates that for each additional mph of wind speed, the expected electricity production increases by 0.240 amperes. Thus, the expected electricity production is $10 \times 0.240 = 2.40$ amperes higher on a day with 25 mph wind velocity as compared to a day with 15 mph wind velocity.

Essentially correct (E) if the response identifies and uses the correct slope value (0.240) *OR* the slope value identified in part (a) of the response

AND

the response includes the following three components:

1. Shows work (correct multiplication or correct substitution into an appropriate expression).
2. Arrives at an answer.
3. Provides correct measurement units (amperes).

Note: Calculating predicted values for both wind speeds and taking their difference is sufficient, as long as measurement units are provided.

Partially correct (P) if the response identifies and uses the correct slope value (0.240) or the slope value identified in part (a) of the response *AND* includes exactly two of the three components listed above.

Incorrect (I) if the response does not meet the criteria for E or P.

Predictor	Coef	SE Coef	T	P
Constant	0.137	0.126	1.09	0.289
Wind velocity	0.240	0.019	12.63	0.000

S = 0.237 R-Sq = 0.873 R-Sq (adj) = 0.868

(c) What proportion of the variation in electricity production is explained by its linear relationship with wind velocity?

The proportion of variation in electricity production that is explained by the linear relationship with wind speed is R^2, which the regression output reports to be 0.873.

Essentially correct (E) if response is 0.873.

Note: No work needs to be shown to earn an E, because the answer is read from the computer output.

Partially correct (P) if the response gives the value of adjusted R^2, rather than R^2, *OR* the response approximates (or rounds) the value of R^2.

Incorrect (I) if the response gives neither R^2 nor adjusted R^2, or if the response reports the square root of R^2.

Predictor	Coef	SE Coef	T	P
Constant	0.137	0.126	1.09	0.289
Wind velocity	0.240	0.019	12.63	0.000

S = 0.237 R-Sq = 0.873 R-Sq (adj) = 0.868

(d) **Is there statistically convincing evidence that electricity production by the windmill is related to wind velocity? Explain.**

Yes, there is very strong statistical evidence that the population slope differs from zero, so electricity production is linearly related to wind speed. For testing the hypotheses $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$, where β represents the population slope, the output reveals that the test statistic is $t = 12.63$ and the p-value (to three decimal places) is 0.000. Because the p-value is so small (much less than both 0.05 and 0.01), the sample data provide very strong statistical evidence that electricity production is linearly related to wind speed.

Essentially correct (E) if the response includes the following three components:

1. Gives the correct conclusion based on a test for the population slope.
2. Reports the correct p-value and/or t-statistic.
3. Provides linkage/justification between the p-value (or t-statistic) and the conclusion.

Partially correct (P) if the response provides exactly two of the three components listed above.

Note: If the wrong p-value is chosen, but the conclusion is consistent with that p-value and linkage or justification is provided, the response earns a P.

Incorrect (I) if the response fails to meet the criteria for E or P.

Homework

page 673 (2, 3, 13)