Exam 1

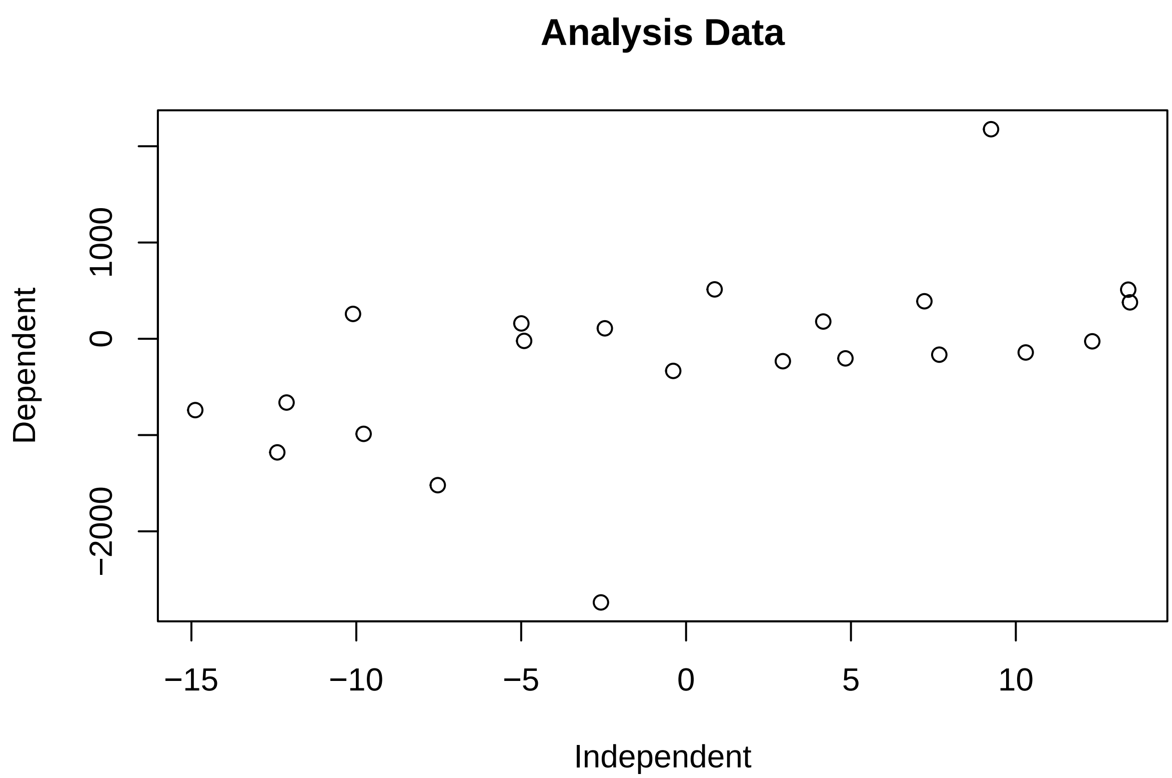
Sebastian Nuxoll

2024-09-16

Initialization

Be sure to enter your name and Vandal number in the YAML header above. Then run the chunk below to create your personalized "analysis_data" for your work. *DO NOT ALTER ANYTHING IN THIS CHUNK.*

Here is what the basic scatter plot of your data looks like:



t table

Here is a table of hypothetical t-values. Use these in the construction of confidence intervals and assignment of p-values. Assume that values are ordered **smallest to largest** when going from **left to right** in the table. The table functions like the `qt()` and `pt()` commands in R. For example, `pt(-a, 10)` would give `0.025` and `pt(a,10)` would give 0.975. Similarly, `qt(0.025,10)` would give `-a`, while `qt(0.975,10)` yields `a`.

Table 1: Hypothetical t-values

df	P = 0.025	P = 0.05	P = 0.1	P = 0.2	P = 0.5	P = 0.8	P = 0.9	P = 0.95	P = 0.975
10	-a	-k	-A	-K	0	K	A	k	a
20	-b	-l	-B	-L	0	L	B	l	b
30	-c	-m	-C	-M	0	M	C	m	c
40	-d	-n	-D	-N	0	N	D	n	d
50	-e	-o	-E	-O	0	O	E	o	e
60	-f	-p	-F	-P	0	P	F	p	f
70	-g	-q	-G	-Q	0	Q	G	q	g
80	-h	-r	-H	-R	0	R	H	r	h
90	-i	-s	-I	-S	0	S	I	s	i
100	-j	-t	-J	-T	0	T	J	t	j

Questions

1. Assume that $\bar{x} = 10$, $\bar{x^2} = 200$, $\bar{y} = 5$, $\bar{y^2} = 50$, and $\bar{xy} = 75$. What are $\beta_0, \beta_1, \sigma_\epsilon, \rho$ and R^2 for the simple regression?

$$\beta_1 = \frac{\bar{xy} - \bar{x}\bar{y}}{\bar{x^2} - \bar{x}^2} = \frac{75 - 10(5)}{200 - 10^2} = \boxed{0.25}$$

$$\beta_0 = \bar{y} - \beta_1\bar{x} = 5 - 0.25(10) = \boxed{2.5}$$

$$\rho = \frac{\bar{xy} - \bar{x}\bar{y}}{\sqrt{(\bar{x^2} - \bar{x}^2)(\bar{y^2} - \bar{y}^2)}} = \frac{75 - 10(5)}{\sqrt{(200 - 10^2)(50 - 5^2)}} = \boxed{0.5}$$

$$R^2 = \rho^2 = 0.5^2 = \boxed{0.25}$$

$$\sigma_\epsilon = \sqrt{(\bar{y^2} - \bar{y}^2)(1 - R^2)} = \sqrt{(50 - 5^2)(1 - 0.25)} \approx \boxed{4.33}$$

- Now assume that $x' = x + 4$ and $y' = y - 2$. Which of the parameters from the previous part changed for the simple regression of y' on x'? What are the new values for those parameters that changed?

The only parameter affected is B_0, which becomes $3 - 0.25(14) = -0.5$

- Lastly, assume that $x'' = 2x$. Again, find which parameters would change and their values for the regression of y on x''.

B_0, B_1, and σ_ϵ are all doubled, making them 5, 0.5, and 8.66 respectively.

2. What is the model underlying simple regression? First write the model in terms of $\mu_i = f(x_i)$. Then specify how y_i relates to μ_i. From that relationship, find the formula for the residual ϵ_i and give the distribution for all residuals.

$$\mu_i = \beta_0 + \beta_1 x_i$$

$$y_i - \mu_i = \epsilon_i \sim N(0, \sigma^2)$$

- How many parameters are we estimating in this simple regression model?

We are estimating 2 parameters: β_0 and β_1.

- Finally, state what the assumptions are from the model you have outlined above.

We assume that there is a linear relationship between x and y and that error is identically distributed across the line.

3. Assume that $\hat{\beta}_1 = 2$ and $\sigma_{\hat{\beta}_1} = 4$ in a model based on 42 observations. Using the table provided above, write the 80% confidence for β_1.

$$\hat{\beta}_1 \pm t_{\alpha/2}\sigma_{\hat{\beta}_1} = \boxed{2 \pm 4D}$$

- Next, assume we want to test $H_0 : \beta \leq 5$, $H_a : \beta > 5$ at the $\alpha = 0.05$ uncertainty level. Write the inequality that we would use to decide whether or not to reject H_0. (In other words, if the inequality is **true** you would reject H_0.)

$$\frac{\hat{\beta}_1 - 5}{\sigma_{\hat{\beta}_1}} > t_{0.05} \Rightarrow \boxed{-0.75 > n}$$

- Finally, assume we want to test $H_0 : \beta = 1$, $H_a : \beta \neq 1$ at the $\alpha = 0.1$ uncertainty level. What is the inequality that we would use to decide whether or not to reject H_0?

$$\left| \frac{\hat{\beta}_1 - 1}{\sigma_{\hat{\beta}_1}} \right| > t_{0.05} \Rightarrow \boxed{0.25 > n}$$

4. Given that $\hat{\beta}_0 = 2, \hat{\beta}_1 = 3, x = 2, n = 82, \sigma_\epsilon = 3,$ and $\sigma_{\hat{y}}(x = 2) = 4$. What is the 60% **confidence** interval at $x = 2$? (Use the hypothetical t table to find the correct values.)

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{\alpha/2}\sigma_{\hat{y}}(x) = 2 + 3(2) \pm 4R = \boxed{8 \pm 4R}$$

- What is the 95% confidence interval for a **prediction** at $x = 2$?

$$\hat{y} \pm t_{\alpha/2}\sqrt{\sigma_{\hat{y}}^2 + \sigma_\epsilon^2} = 8 \pm r\sqrt{4^2 + 9^2} = \boxed{8 \pm \sqrt{97}r}$$

5. Use `lm()` to perform the simple regression of the dependent variable on the independent variable in the `analysis_data` tibble and answer the following the questions:

- What is the fraction of variance that is *unexplained* by the model?
- What is the slope of the fitted line?
- What is the probability that β_0 is 0?
- Assume that we are interested in whether the slope is greater than 10. Calculate the appropriate t-value. Write the R command that finds the p-value associated with that t-value and the appropriate hypothesis.
- Assume that we want to demonstrate that the intercept is not -5. Calculate the appropriate t-value. Write the R command that finds the p-value associated with that t-value and the appropriate hypothesis.
- Run another regression that hypothesizes that the dependent variable is related to the **squared** independent variable. Compare the two models and give an argument as to which model is the best (support your argument with values from the regressions).

6. Use **ggplot** to create the following the plots:

- A plot that has the raw data and the first fitted line from (5).
- A plot that has the raw data, the first fitted line from (5), and the *confidence interval* for the line assuming that you want the interval associated with $t = \pm 1.3$.
- A plot that has the raw data, the first fitted line from (5), and the *prediction interval* for the line assuming that you want the interval associated with $t = \pm 2$.
- A plot that has the raw data and *both* of the fitted models from (5).

7. Use `dplyr` commands to do the following:

- Sort the analysis data based on the independent variable
- Create a new column called “AB” that is a factor where 50% of values are “A” and the remainder are “B”.
- Group the variables by “AB” and find the mean values of the independent and dependent variables using the `summarize()` command.
- Drop all observations from the data that are in the lower quartile of the dependent variable.
- Create a new column called “Transform” that contains a mathematical transform of the dependent and independent variables (you can choose whatever function you want).
- Create a new tibble that only retains the “AB” and the “Transform” columns.
- Print out the summary of this newest tibble.