Regression Problem Examples

Shuangjie Zhang

Feb 2020

Please include explanations in the homework and exam, they accounts for most points. (Each step + why do this step).

1 Estimate/ Predict the average

Questions looks like: Given average of two variables, SD of two variables, and correlation r. We are asked to **estimate/predict the average value of prediction variable when given a new given value**.

We have two ways to calculate the prediction number.

1.1 Z score method (CH 10)

(0) Identify which is given a number, it is the given variable. What we need to predict is prediction variable.

(1) Calculate the Z score for given variable.

$$Z_{\text{Given}} = \frac{\text{Given value} - \text{Avg of Given Variable}}{\text{SD of Given Variable}}$$

(2) Calculate the Z score for prediction variable.

$$Z_{\text{Prediction}} = r \times Z_{\text{Given}}$$

(3) Calculate the prediction value.

$$\text{Prediction Value} = \text{Avg}_{\text{Prediction Variable}} + Z_{\text{Prediction}} \times SD_{\text{Prediction Variable}}$$

1.2 Slope + Intercept (CH 12)

(1) Calculate the slope for regression line.

$$\text{slope} = r \times \frac{\text{SD of Prediction Variable}}{\text{SD of Given Variable}}$$

(2) Calculate the intercept.

$$\text{intercept} = \text{Avg of Prediction Variable} - \text{slope} \times \text{Avg of Given Variable}$$

(3) Predict the prediction value

$$\text{Prediction Value} = \text{intercept} + \text{slope} \times \text{Given Value}$$

1.3 Example

Average age = 50 years, SD of age = 16 years. Average educational level = 13.2 years, SD of educational level = 3.0 years, r = -0.20.

(1) Predict the average educational level when age is 60 years.

(2) Predict the average age when educational level is 16.2 years.

Solution:

(1) Given variable is age and prediction variable is educational level.

Z score method:

Z score for age: $Z_{\text{age}} = \frac{60-50}{16} = 0.625$

Z score for educational level: $Z = r \times Z_{\text{age}} = -0.20 \times 0.625 = -0.125$

The average educational level when age is 60 years:

$y = 13.2 + (-0.125) \times 3.0 = 12.825$ years

Slope + intercept method:

Slope $= r \times \frac{\text{SD of educational level}}{\text{SD of age}} = -0.20 \times \frac{3}{16} = -0.0375$

Intercept = Avg of y - slope × Avg of x = 13.2 - (- 0.0375) × 50 = 15.075

Write down the regression line formula: $y = 15.075 - 0.00375 \times x$

The average educational level when age is 60 years:

$y = 15.075 - 0.00375 \times 60 = 12.825$ years

(2) Given variable is educational level and prediction variable is age.

Z score method:
Z score for educational level: $Z_{\text{educational level}} = \frac{16.2-13.2}{3} = 1$
Z score for age: $Z = r \times Z_{\text{educational level}} = -0.20 \times 1 = -0.20$
The average age when educational level is 16.2 years:
$y = 50 + (-0.2) \times 16 = 46.8$ years

Slope + intercept method:
Slope $= r \times \frac{\text{SD of age}}{\text{SD of educational level}} = -0.20 \times \frac{16}{3} = -1.06666666667$
Intercept = Avg of y - slope × Avg of x = 50 - (-1.06666666667) × 13.2 = 64.08
Write down the regression line formula: $y = 64.08 - 1.06666666667 \times x$
The average age when educational level is 16.2 years:
$y = 64.08 - 1.06666666667 \times 16.2 = 46.8$ years

2 Percentile in regression (CH 10 Q 9)

Questions looks like: Given correlation r between given variable and prediction variable. We are asked to **predict the percentile rank in prediction variable for a given percentile rank in given variable**.

2.1 Steps for predicting the percentile rank

(0) Identify given variable and prediction variable. Variable which is given a given percentile rank is given variable. The other one is prediction variable.

(1) Calculate Z score for given variable.

Use chapter 5 knowledge first to calculate the **middle area** using symmetry property of normal. Draw a picture helps you a lot.

More explicitly, if you are comfortable with math formula, here is a direct way to calculate the middle area. Suppose we are given α th percentile. If $\alpha < 50$, the middle area will be $2 * (50 - \alpha)\%$. And if $\alpha > 50$, the middle area will be $2 * (\alpha - 50)\%$. In summary, the middle area will be$2 * |\alpha - 50|\%$

After we get the **middle area**, from the normal table, we will get the Z score for given variable.

(2) Calculate Z score for prediction variable.

$$Z_{\text{Prediction}} = r \times Z_{\text{Given}}$$

(3) Calculate middle area for prediction variable.

Because of symmetry, and also refer to chapter 5 knowledge again. Use normal table to find the middle area for y. And we need to identify whether use 50% minus half of the middle are or use 50% plus half of the middle area. Drawing a picture helps a lot. The key is whether $Z_{\text{Prediction}}$ is positive or negative. If $Z_{\text{Prediction}}$ is positive, we use 50% plus half of the middle area. If $Z_{\text{Prediction}}$ is negative, we use 50% plus half of the middle area.

2.2 Example

Average SAT score = 550, SD = 80. Average first year GPA = 2.6, SD = 0.6, r= 0.40. The scatter plot is foot-ball shaped.

(1)Suppose the percentile rank of one student on the SAT is 90th, among the first-year students. Predict his percentile rank on first-year GPA. The scatter plot is foot-ball shaped.

(2)Suppose the percentile rank of one student on the SAT is 80th, among the first-year students. Predict his percentile rank on first-year GPA. The scatter plot is foot-ball shaped.

(3)Suppose we have changed the correlation from $r = 0.40$ to $r = -0.40$. Again, suppose the percentile rank of one student on the SAT is 90th, among the first-year students. Predict his percentile rank on first-year GPA.

Solution:

(1) Given variable is SAT score and prediction variable is first-year GPA.

[**Step 1**] Calculate the middle area for given variable and find the Z score for given variable.

Since 90th percentile means area below this number is 90%, the rest right area will be 10%. Because of **symmetry**, the middle area will be: 100%-2× 10%=80%. From the normal table, we find that $Z_{\text{Given}} = 1.30$.

[**Step 2**] Calculate the Z score for prediction variable.

$Z_{\text{Prediction}} = r \times Z_{\text{Given}} = 0.4 \times 1.30 = 0.52$

[**Step 3**] Find the middle area for prediction variable and find the percentile rank for prediction variable.

From normal table, when $z = 0.52$ is close to 0.5, the middle are is 38.29%. Because of symmetry, half of middle area is $\frac{38.29\%}{2} = 19.145\%$. Draw a picture, then we will see that this percentile rank is above the average. SO we use 50 plus half of middle area. $50\% + \frac{38.29\%}{2} = 50\% + 19.145\% = 69.145\%$. So the percentile rank on first-year GPA is 69.145th.

(2) Given variable is SAT score and prediction variable is first-year GPA.

[**Step 1**] Calculate the middle area for given variable and find the Z score for given variable.

Since 80th percentile means area below this number is 80%, the rest right area will be 20%. Because of **symmetry**, the middle area will be: 100%-2× 20%=60%. From the normal table, we find that $Z_{\text{Given}} = 0.85$.

[**Step 2**] Calculate the Z score for prediction variable.

$Z_{\text{Prediction}} = r \times Z_{\text{Given}} = 0.4 \times 0.85 = 0.34$

[**Step 3**] Find the middle area for prediction variable and find the percentile rank for prediction variable.

From normal table, when $z = 0.34$ is close to 0.35, the middle are is 27.37%. Because of symmetry, half of middle area is $\frac{27.37\%}{2} = 13.685\%$. Draw a picture, then we will see that this percentile rank is above the average. SO we use 50 plus half of middle area. $50\% + \frac{27.37\%}{2} = 50\% + 13.685\% = 63.685\%$. So the percentile rank on first-year GPA is 63.685th.

(3) Given variable is SAT score and prediction variable is first-year GPA.

[**Step 1**] Calculate the middle area for given variable and find the Z score for given variable.

Since 90th percentile means area below this number is 90%, the rest right area will be 10%. Because of **symmetry**, the middle area will be: 100%-2× 10%=80%. From the normal table, we find that $Z_{\text{Given}} = 1.30$.

[**Step 2**] Calculate the Z score for prediction variable.

$Z_{\text{Prediction}} = r \times Z_{\text{Given}} = \mathbf{-0.4} \times 1.30 = \mathbf{-0.52}$

[**Step 3**] Find the middle area for prediction variable and find the percentile rank for prediction variable.

From normal table, when $z = -0.52$, $|z| = 0.52$ is close to 0.5, the middle are between [-0.52, 0.52] is 38.29%. Because of symmetry, half of middle area is $\frac{38.29\%}{2} = 19.145\%$. Draw a picture. Since$Z_{\text{Prediction}} < 0$, we use 50% minus half of middle area. $50\% - \frac{38.29\%}{2} = 50\% - 19.145\% = 30.855\%$. So the percentile rank on first-year GPA is 30.855th.

3 R.M.S error related new prediction problem (CH11 Q 5)

Questions looks like: Given average of two variables, SD of two variables,and correlation r. We are asked: **Among those who have given value on given variable, what percentage has prediction variable above(below) some number.**

The key point here is for prediction:
New average = Prediction Value
New SD = R.M.S error

3.1 Steps for solving this kind of question

(0) Identify given variable and prediction variable.

Variable which is given a certain number not a region is given variable. Variable which is asked with a region above(below) some number is prediction variable.

(1) Calculate new average.

That is exactly the prediction value. Both Z score method and slop+intercept method can be applied here.

(2) Calculate new SD. That is equal to R.M.S error.

$$\text{R.M.S error} = \sqrt{1-r^2} \times SD_{\text{Prediction}}$$

(3) Find the area has been asked in the question.

Use chapter 5 knowledge again. First standardize the area boundary number. Get the new Z score. Then, refer to normal table and get the middle area. Use 50% plus/minus half of middle area will be the answer.

3.2 Example

Average LSAT score = 162, SD = 6. Average first year score = 68, SD = 10, r= 0.60.

Q: Among the student who scored 174 on the LSAT, about what percentage had first year scores over 86?

Solution:

(0) Identify given variable and prediction variable.

LSAT is given a certain number 174. So LSAT is given variable. First year score is asked with a region above 86. So first year score is prediction variable.

(1) Calculate new average.

Here I use slop+intercept method.

Slope = r $\times \frac{\text{SD of LSAT}}{\text{SD of first year score}} = 0.60 \times \frac{10}{6} = 1$.

Intercept = Avg of first year score - slope × Avg of LSAT = 68-1×162= -94.

Write down regression line formula: y=-94+x

When LSAT = 174, prediction of first year score is 80.

So new average is 80.

(2) Calculate new SD. That is equal to R.M.S error.

$$\text{R.M.S error} = \sqrt{1-r^2} \times SD_{\text{first year score}}$$

$$\text{R.M.S error} = \sqrt{1-0.6^2} \times 10 = 8$$

So new SD is 8.

(3) Find the area has been asked in the question.

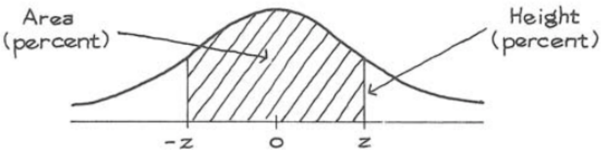
For the new normal, new average is 80. New SD is 8. We are being asked to compute how many percentage over 86. This is again going back to chapter 5 knowledge.

Standardization: $Z = \frac{86-80}{8} = \frac{6}{8} = 0.75$

From the normal table, we find that when z=0.75, the middle area is 54.67%. Since we need to compute how many percentage above 86, we need to use $50\% - \frac{54.67\%}{2} = 22.665\%$.

So the final answer is 22.665%, among the student who scored 174 on the LSAT, about 22.665% had first year scores over 86

Tables



A NORMAL TABLE

z	*Height*	*Area*	*z*	*Height*	*Area*	*z*	*Height*	*Area*
0.00	39.89	0	1.50	12.95	86.64	3.00	0.443	99.730
0.05	39.84	3.99	1.55	12.00	87.89	3.05	0.381	99.771
0.10	39.69	7.97	1.60	11.09	89.04	3.10	0.327	99.806
0.15	39.45	11.92	1.65	10.23	90.11	3.15	0.279	99.837
0.20	39.10	15.85	1.70	9.40	91.09	3.20	0.238	99.863
0.25	38.67	19.74	1.75	8.63	91.99	3.25	0.203	99.885
0.30	38.14	23.58	1.80	7.90	92.81	3.30	0.172	99.903
0.35	37.52	27.37	1.85	7.21	93.57	3.35	0.146	99.919
0.40	36.83	31.08	1.90	6.56	94.26	3.40	0.123	99.933
0.45	36.05	34.73	1.95	5.96	94.88	3.45	0.104	99.944
0.50	35.21	38.29	2.00	5.40	95.45	3.50	0.087	99.953
0.55	34.29	41.77	2.05	4.88	95.96	3.55	0.073	99.961
0.60	33.32	45.15	2.10	4.40	96.43	3.60	0.061	99.968
0.65	32.30	48.43	2.15	3.96	96.84	3.65	0.051	99.974
0.70	31.23	51.61	2.20	3.55	97.22	3.70	0.042	99.978
0.75	30.11	54.67	2.25	3.17	97.56	3.75	0.035	99.982
0.80	28.97	57.63	2.30	2.83	97.86	3.80	0.029	99.986
0.85	27.80	60.47	2.35	2.52	98.12	3.85	0.024	99.988
0.90	26.61	63.19	2.40	2.24	98.36	3.90	0.020	99.990
0.95	25.41	65.79	2.45	1.98	98.57	3.95	0.016	99.992
1.00	24.20	68.27	2.50	1.75	98.76	4.00	0.013	99.9937
1.05	22.99	70.63	2.55	1.54	98.92	4.05	0.011	99.9949
1.10	21.79	72.87	2.60	1.36	99.07	4.10	0.009	99.9959
1.15	20.59	74.99	2.65	1.19	99.20	4.15	0.007	99.9967
1.20	19.42	76.99	2.70	1.04	99.31	4.20	0.006	99.9973
1.25	18.26	78.87	2.75	0.91	99.40	4.25	0.005	99.9979
1.30	17.14	80.64	2.80	0.79	99.49	4.30	0.004	99.9983
1.35	16.04	82.30	2.85	9.69	99.56	4.35	0.003	99.9986
1.40	14.97	83.85	2.90	0.60	99.63	4.40	0.002	99.9989
1.45	13.94	85.29	2.95	0.51	99.68	4.45	0.002	99.9991