STATISTICS AND R LECTURE 6

PETER LAKNER SIMPLE LINEAR REGRESSION

In a simple linear regression model we are interested in the linear relationship between two variables.

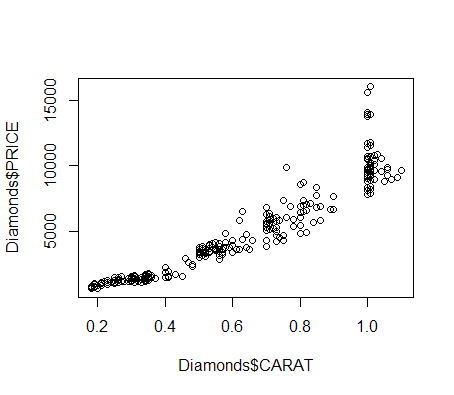
y: Dependent or response variable, for example sale price of an apartment.

x: Independent or predictor variable, for example size of an apartment.

In the next example the predictor will be the size of a diamond (in karats), and he dependent variable will be the asking price (dollars). Click on “Import Dataset” in R, and select the file “Diamonds”.

In order to have a scatterplot, type in R plot(Diamonds$CARAT,Diamonds$PRICE) We receive the following illustration:

1



There seems to be a nice approximate linear relationship between CARAT and PRICE. The strength of the linear relationship is measured by r, the correlation coefficient. Here is a formula for calculating it. Suppose that we have the paired data (x1, y1 ), . . . (xn, yn). Then the correlation coefficient is calculated as

n

(xi − x¯)(yi − y¯)

i=1

r = .

n

2 n

i=1 (xi − x¯)

2

i=1 (yi − y¯)

r is always between -1 and 1. Rule of thumb:

.8 to 1: strong direct linear relationship

.5 to .8: moderate direct linear relationship

.2 to .5: weak direct linear relationship

-.2 to .2: hardly any linear relationship

-.5 to -.2: weak inverse linear relationship

-.8 to -.5: moderate inverse linear relationship

-1 to -.8: strong inverse linear relationship

Calculate r in the case ov CARAT versus PRICE. Type in R

cor(Diamonds$CARAT, Diamonds$PRICE)

We get r = .94, there is a strong direct linear relationship between CARAT and

PRICE.

In a simple linear regression we set up the following model:

yi = β0 + β1xi + Ei

Here β0 and β1 are unknown constants. β0 is called the intercept and β1 is called the slope. Each of the Ei follows normal distribution with mean zero and the same (unknown) variance σ2. We also assume that these E’s are independent.

Our first objective is to find an estimate for β0 and β1 based on the data. We shall use again the Diamonds example. Type in R

regdiamonds=lm(Diamonds$PRICE ∼ Diamonds$CARAT)

regdiamonds

R will return estimates the coefficients βˆ0 and βˆ1.

βˆ0 = −2, 298

βˆ1 = 11, 599

How do we get these estimates for the coefficients? For an arbitrary line let e1 , . . . , en be the vertical distances from the points on the scatter-plot and the line, as shown on the graph below. The regression line is the one that minimizes the sum of the

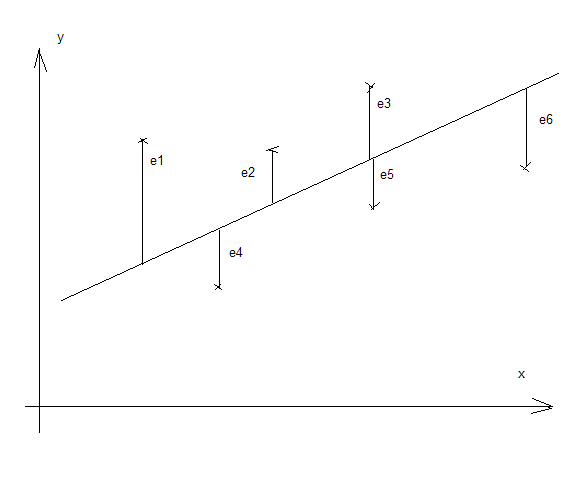
squares n

i=1

e2. The intercept and the slope of the line that minimizes the above

sum of squares will be βˆ0 and βˆ1 .

i



Knowing βˆ0 and

βˆ1 allows to predict the sale price of a diamond. Suppose that a

piece of diamond has x = .4 carat. Then the estimated sale price will be

yˆ = βˆ0 + βˆ1x = −2298 + 11, 599 × .4 = 2, 341.16

Interpretation of βˆ1: In general terms, for every unit increase of the predictor x, the response y is estimated to increase by βˆ1 unit. In the diamond example, every karat increase of the weight of a diamond, the sale price is estimated to increase by 11,599

dollars. Since in practice the weight increases rather by .1 karat, it is more practical to say that for every 0.1 karat increase of the weight of a diamond, the sale price is estimated to increase by 1,159.90 dollars.

Interpretation of

βˆ0: This represents the predicted value when x = 0. In the dia-

mond example zero weight does not make sense, so the intercept has no meaningful interpretation. Also, the intercept turns out to be negative, whereas a price must be positive.

EXAMPLE: We use a file called FCAT. This file contains information pertaining var- ious public schools. To look at a file, click on “Import File” in R, select FCAT.

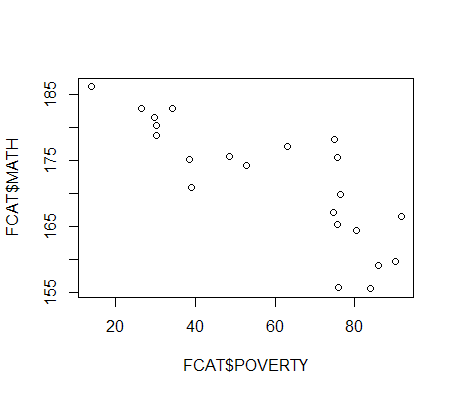
The predictor is POVERTY (this is the percentage of students at each school below poverty level). The response MATH (average result of a standardized math test in each school).

a) Create a scatterplot.

b) Run a regression, find the estimated coefficients. c) Interpret the slope. Interpret the intercept.

d) Estimate the average math test score in a school with POVERTY=20.

SOLUTION. a) Type in R plot(FCAT$POVERTY,FCAT$MATH) We get the following illustration:



b) Type in R

lm(FCAT$MATH∼FCAT$POVERTY)

The results are βˆ0 = 189.8158 and βˆ1 = −0.3054.

c) Interpretation of

βˆ0 = 189.8: the average math score for schools with zero per-

centage of students below poverty level is 189.8. Interpretation of βˆ1 = −.3: as the

percentage of students under poverty level increases by 1 percent, the average math

score decreases by .3 points.

d) Here is the estimate:

yˆ = 189.8 − .3 × 20 = 183.8

Based on the linear model, we expect the average score for schools with P OV ERT Y =

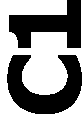
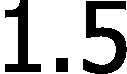
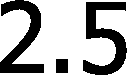
20 to be 183.8.

Do not try to estimate the response for x values which are too much “off the data- base”. Let’ look at the “Diamonds” file again. Type in R summary(Diamonds$CARAT)

We get minimum= .18 and maximum=1.1. The estimation will not work well if we go very much off this range, say, for CARAT= .03 or for CARAT= 3.5.

INFERENCES ABOUT THE SLOPE. If β1 = 0 then the regression model does not explain the variation of the dependent variable. On the scatter-plot below the slope is not exactly zero but close to it. Notice that in that case the variation of x has very small impact on y.

For this reason it is important to test whether the slope is zero or not. We set up two hypotheses: H0 : β1 = 0 and H1 : β1 /= 0.



Type the following in R:

regfcat=lm(FCAT$MATH∼FCAT$POVERTY)

summary(regfcat)

R will return the following information:

Call:

lm(formula = FCAT$MATH ∼FCAT$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 \*\*\* |
| FCAT$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

We have a lot of information here, but at this point we are only interested in the line headed with “FCAT$POVERTY”. First the value of βˆ1 = −.3 given. Next, in

the “Std. Error” column we see the figure .05 (rounded from .04759). This is the estimated standard deviation of βˆ1, denoted by sβˆ . This number gives an idea how good an estimate βˆ1 is. For example, we can create a 95% confidence interval for the true slope using the formula

1

βˆ1 ± t.025 × sβˆ

1

where the degree of freedom for the t.025 here is n − 2 (n is the size of the columns in

the data set). In our case n = 22. We type in R

qt(.975,df=20)

and see that t.025 = 2.08. So the 95% confidence interval is −.3 ± 2.08 × .05, that is,

(−.4, −.2) (rounded).

We can also test the two hypothesis listed above. The tstat is calculated as

βˆ1

tstat = ,

s ˆ

β1

but the value is printed there in the “t value” column, it is tstat = −6.418. We select the significance level α = .05, in which case we reject the H0 if the tstat value is outside

of the interval (−t.025, t.025). In this case -6.4 is indeed does not fall in the interval

(−2.08, 2.08), so we reject H0 .

R can construct a confidence interval for the real slope β1 saving as the work. We select the confidence level .95, then type in R

confint(regfcat, ’FCAT$POVERTY’, level=0.95)

and we get the same result as above, i.e., (−.4, −.2). (If copy/paste from the PDF file

to R then the quotation marks need to be re-typed in R.) Remember that we already created a variable “regfcat” above, in which the linear model for the FCAT file is

stored. We are 95% confident that the real slope β1 falls in the interval (−.4, −.2).

What is the 2.93e-06 figure in the last column? This is the p-value of the test. There is a really fast way of testing the hypothesis H0 : β1 = 0 versus H1 : β1 /= 0. Compare

the p-value to the significance level. If the p-value is below the significance level, we reject H0 ; otherwise we can not reject it. In our case the p-value is tiny, certainly below .05, so we can reject the null-hypothesis (we already knew that, of course).

Meaning of the p-value: how likely is to receive such data under the null-hypothesis. In our case it is extremely unlikely, so we reject the H0 .

The figure .6731 next to the words “Multiple R-squared” is called the coefficient of determination. The coefficient of determination represents the proportion of the to- tal variation of the observations around the mean y¯ that is explained by the linear relationship between x and y. In our case, 67% of the variation of the MATH score is explained by the proportion of students living below poverty level.

The coefficient of determination is always between 0 and 1. It is calculated the same way and has the same interpretation in the multiple regression case (when there are several independent variables). However, in the simple regression it has an additional interpretation: it is exactly the square of the correlation coefficient between x and y. Larger r2 value indicates that the linear regression model has more explaining power. Here is a rule of thumb:

.65 ≤ r2 ≤ 1: strong model;

.25 ≤ r2 < .65: the model has moderate strength;

0 ≤ r2 < .25: the model is hardly worth considering in its present form.

In our case the linear regression model is strong.

USING LINEAR REGRESSION FOR ESTIMATION AND PREDICTION.

We shall use the “Diamonds” file as our example. Let us suppose that we want to estimate the sale price of a diamond with karat=.6.

We already know that βˆ0 = −2, 298 and βˆ1 = 11, 599. Thus our estimate is

yˆ = −2, 298 + 11, 599 × .6 = 4, 661

Next we shall create a confidence interval for the average price of 0.6 karat diamonds. Type in R:

x=Diamonds$CARAT

y=Diamonds$PRICE

new = data.frame(x = .6)

conf = predict(lm(y∼x), new, interval=”confidence”)

conf

We get a 95% confidence interval for the average price of 0.6 carat diamonds, it is

(4, 534.889, 4787.056). Of course we shall round this to (4, 535, 4787).

Next we shall find a 95% prediction interval for the price of an individual xp = .6 karat diamond. Type in R:

pred = predict(lm(y ∼ x), new, interval=”prediction”)

pred

We get the (rounded) prediction interval (2, 458, 6, 864). Notice that the prediction interval is much larger than the confidence interval. It is much more difficult to pre- dict the price if an individual piece of diamond than to estimate the average price of diamonds with a given weight (CARAT).

The values yˆi = βˆ0 + βˆ1 xi are called fitted values.

Here is a graphical representation of the prediction and confidence intervals. Type in

R the following:

pred=predict(regdiamonds, interval=”prediction”)

confid=predict(regdiamonds, interval=”confidence”)

R will give a “warning message”, but it can be safely disregarded.

This will create matrices called “pred” and “confid”. The “pred” matrix will have

3 columns, the fitted values, the lower prediction limits, and the upper prediction limits. The “confid” matrix will have the same but with confidence limits.

plot(x,y)

lines(x,pred[,1], col=”blue”,lwd=2) lines(x,pred[,2], col=”green”,lwd=2) lines(x,pred[,3], col=”green”,lwd=2) lines(x,confid[,2], col=”red”,lwd=2) lines(x,confid[,3], col=”red”,lwd=2)

