Simple Linear Regression: Inference, Prediction, Explanation

Christian Thieme

2/17/2021

A Modern Approach to Regression with R

Exercise 2.3

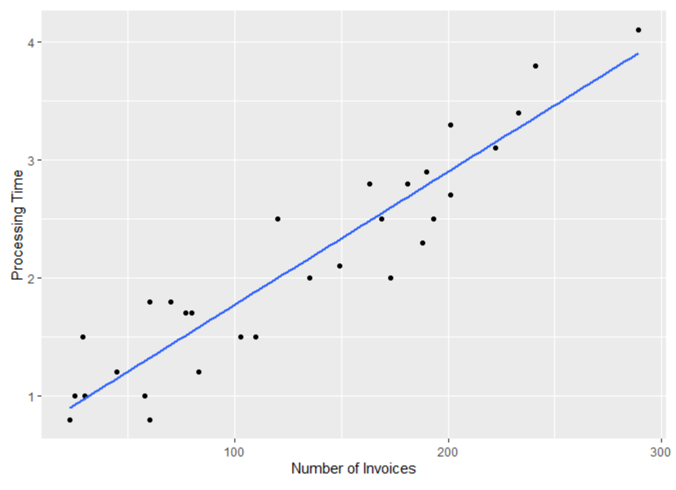
The manager of the purchasing department of a large company would like to develop a regression model to predict the average amount of time it takes to process a given number of invoices. Over a 30-day period, data are collected on the number of invoices processed and the total time taken (in hours). The data are available on the book web site in the file invoices.txt. The following model was fit to the data: $Y = \beta_0 + \beta_1 + e$ where Y is the processing time and x is the number of invoices. A plot of the data and the fitted model can be found in Figure 2.7. Utilizing the output from the fit of this model provided below, complete the following tasks.

- **a. Find a 95% confidence interval for the start-up time, i.e., β_0.**

Here we will be looking to build a 95% confidence interval for the Y-intercept.

```
invoices <- read_tsv('https://gattonweb.uky.edu/sheather/book/docs/datasets/invoices.txt')
```

```
ggplot(invoices) +
  aes(x = Invoices, y = Time) +
  geom_point() +
  ylab("Processing Time") +
  xlab("Number of Invoices") +
  geom_smooth(method = lm, se = FALSE)
```




```
model <- lm(Time ~ Invoices, data = invoices)

summary(model)
```

```
##
## Call:
## lm(formula = Time ~ Invoices, data = invoices)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59516 -0.27851  0.03485  0.19346  0.53083
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.6417099  0.1222707   5.248 1.41e-05 ***
## Invoices    0.0112916  0.0008184  13.797 5.17e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3298 on 28 degrees of freedom
## Multiple R-squared:  0.8718, Adjusted R-squared:  0.8672
## F-statistic: 190.4 on 1 and 28 DF,  p-value: 5.175e-14
```

We can use the output above to calculate the confidence interval by hand:

```
t_value <- qt(0.975, df = 28)
se <- 0.1222707

0.6417099 + c(-1,1) * t_value * se
```

```
## [1] 0.3912497 0.8921701
```

We can see that the confidence interval above does not include zero, so we would reject the null hypothesis that the intercept is equal to 0 (if we were running that test). Alternately, we can use the `confint` function to calculate the same values.

```
confint(model, level = 0.95)[1,]
```

```
##     2.5 %    97.5 %
## 0.3912496 0.8921701
```

The 95% confidence interval for β_0 is (0.3912496,0.8921701). This means that we can expect to find the actual Y-intercept within this range 95% of the time if we drew random samples. As you can see, it is a very wide range. We would expect that as we only had 30 samples in our dataset. The more samples we include, the more narrow our confidence interval would be.

- **b. Suppose that a best practice benchmark for the average processing time for an additional invoice is 0.01 hours (or 0.6 minutes). Test the null hypothesis $H_0 : \beta_1 = 0.01$ against a two-sided alternative. Interpret your result.**

Similar to above, we can calculate the the confidence interval for a given confidence. For our first test, we'll use a 95% confidence interval:

```
t_value <- qt(0.975, df = 28)
se <- 0.0008184

0.0112916 + c(-1,1) * t_value * se
```

```
## [1] 0.009615184 0.012968016
```

The value 0.01 falls within the confidence interval, so we would fail to reject the null hypothesis and say that there is no significant evidence that the average processing time is different than the benchmark. Now, let's see if this result changes if we move to a 99% confidence interval:

```
t_value <- qt(0.995, df = 28)
se <- 0.0008184

0.0112916 + c(-1,1) * t_value * se
```

```
## [1] 0.009030146 0.013553054
```

It looks like our value of 0.01 still falls within our confidence interval, so our previous conclusion would not change.

- **c. Find a point estimate and a 95% prediction interval for the time taken to process 130 invoices.**

To solve this manually, we can use the estimates from the model output above:

```
intercept <- 0.6417099
invoice_slope <- 0.0112916
invoice_num <- 130


point_estimate <- intercept + (invoice_slope * invoice_num)

df = 28

t_value <- qt(0.975, df = df)
rse <- 0.3298
rss <- rse ^2 * df

point_estimate + c(-1,1) * t_value * se
```

```
## [1] 2.107941 2.111294
```

We can find our prediction point as well as our confidence interval dynamically with this line of code:

```
predict(model, data.frame(Invoices = 130), interval = "prediction")
```

```
##        fit      lwr    upr
## 1 2.109624 1.422947 2.7963
```

Linear Models with R

Exercise 3.4

Using the `sat` data:

- **a. Fit a model with `total` sat score as the response and `expend`, `ratio` and `salary` as predictors. Test the hypothesis that $\beta_{salary} = 0$. Test the hypothesis that $\beta_{salary} = \beta_{ratio} = \beta_{expend} = 0$. Do any of these predictors have an effect on the response?**

```
head(sat)
```

```
##            expend ratio salary takers verbal math total
## Alabama     4.405  17.2 31.144      8    491  538  1029
## Alaska      8.963  17.6 47.951     47    445  489   934
## Arizona     4.778  19.3 32.175     27    448  496   944
## Arkansas    4.459  17.1 28.934      6    482  523  1005
## California  4.992  24.0 41.078     45    417  485   902
## Colorado    5.443  18.4 34.571     29    462  518   980
```

First, we'll test the hypothesis that $\beta_{salary} = 0$. In order to test this hypothesis, we'll initialize a model with all of the predictors.

```
model1 <- lm(total ~ expend + ratio + salary, data = sat)
summary(model1)
```

```
##
## Call:
## lm(formula = total ~ expend + ratio + salary, data = sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -140.911  -46.740   -7.535   47.966  123.329
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1069.234    110.925   9.639 1.29e-12 ***
## expend        16.469     22.050   0.747   0.4589
## ratio          6.330      6.542   0.968   0.3383
## salary        -8.823      4.697  -1.878   0.0667 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.65 on 46 degrees of freedom
## Multiple R-squared:  0.2096, Adjusted R-squared:  0.1581
## F-statistic: 4.066 on 3 and 46 DF,  p-value: 0.01209
```

Next, we'll initialize another model, but this time, we'll remove salary from the model and then run an anova over the data:

```
model2 <- lm(total ~ expend + ratio, data = sat)
anova(model2, model1)
```

```
## Analysis of Variance Table
##
## Model 1: total ~ expend + ratio
## Model 2: total ~ expend + ratio + salary
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     47 233443
## 2     46 216812  1     16631 3.5285 0.06667 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the above output, we can see the p-value of 0.06667 is above 0.05. We will fail to reject the null hypothesis and conclude that there is not significant evidence to say the corresponding parameter for salary is not 0.

Now, let's test the hypothesis that $\beta_{salary} = \beta_{ratio} = \beta_{expend} = 0$. To do this, we'll initialize the null model and use the `anova` function:

```
nullmod <- lm(total ~ 1, data = sat)
anova(nullmod, model1)
```

```
## Analysis of Variance Table
##
## Model 1: total ~ 1
## Model 2: total ~ expend + ratio + salary
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     49 274308
## 2     46 216812  3     57496 4.0662 0.01209 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the p-value above of 0.01209, we will reject the null hypothesis and say that there is evidence that at least one of these coefficients is not 0.

- **b. Now add `takers` to the model. Test the hypothesis that $\beta_{takers} = 0$. Compare this model to the previous one using an F-test. Demonstrate that the F-test and t-test here are equivalent.**

```
model4 <- lm(total ~ expend + ratio + salary + takers, data = sat)

anova(model1, model4)
```

```
## Analysis of Variance Table
##
## Model 1: total ~ expend + ratio + salary
## Model 2: total ~ expend + ratio + salary + takers
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     46 216812
## 2     45  48124  1    168688 157.74 2.607e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the output of the anova, we would reject the null hypothesis that the coefficient for takers is equal to zero.

The F-statistic for this test is 157.74. If we calculate the t-statistic and square it, we will get the F-statistic. The to values from the F-statistic and t-statistic are equal within rounding error.

```
tstat <- (-2.9045 - 0)/0.2313
tstat^2
```

```
## [1] 157.6854
```

```
2*pt(tstat, 45)
```

```
## [1] 2.621879e-16
```