simple_regression

Lana Gruncic Krajnc

2023-07-12

The file COLLIN.txt contains data for 21 110m hurdles runs by runner Collin Jackson:

wind speed = windspeed (m/s) and

running time = time (s).

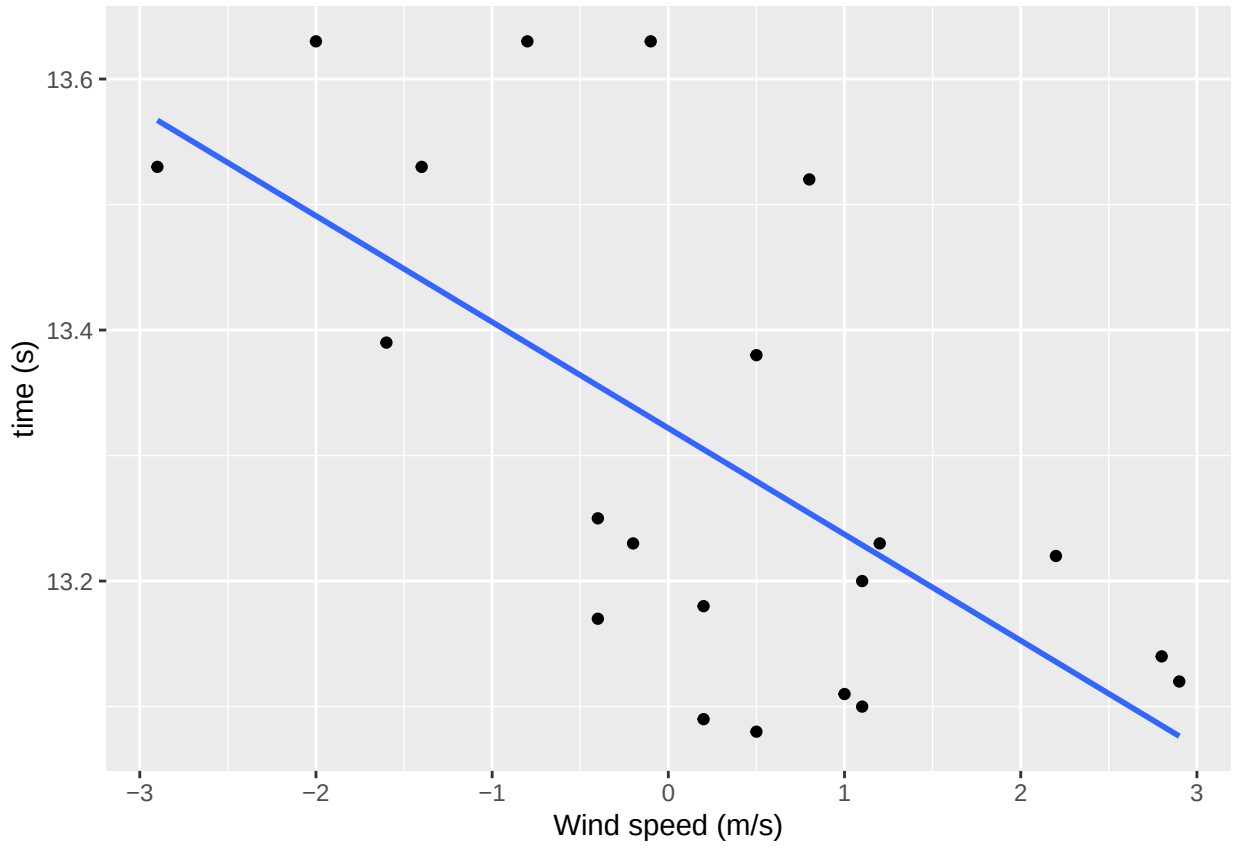
The data were obtained in an experiment in an enclosed space (Source: Daly et al., p. 525). The wind speed was selected in advance for each run separately. Negative wind speed values mean that the wind blew into the runner's chest. We want to explain how wind speed affects 110m running time?

Data presentation

```
## Warning: package 'ggplot2' was built under R version 4.2.2
```

```
## Warning: package 'gridExtra' was built under R version 4.2.2
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

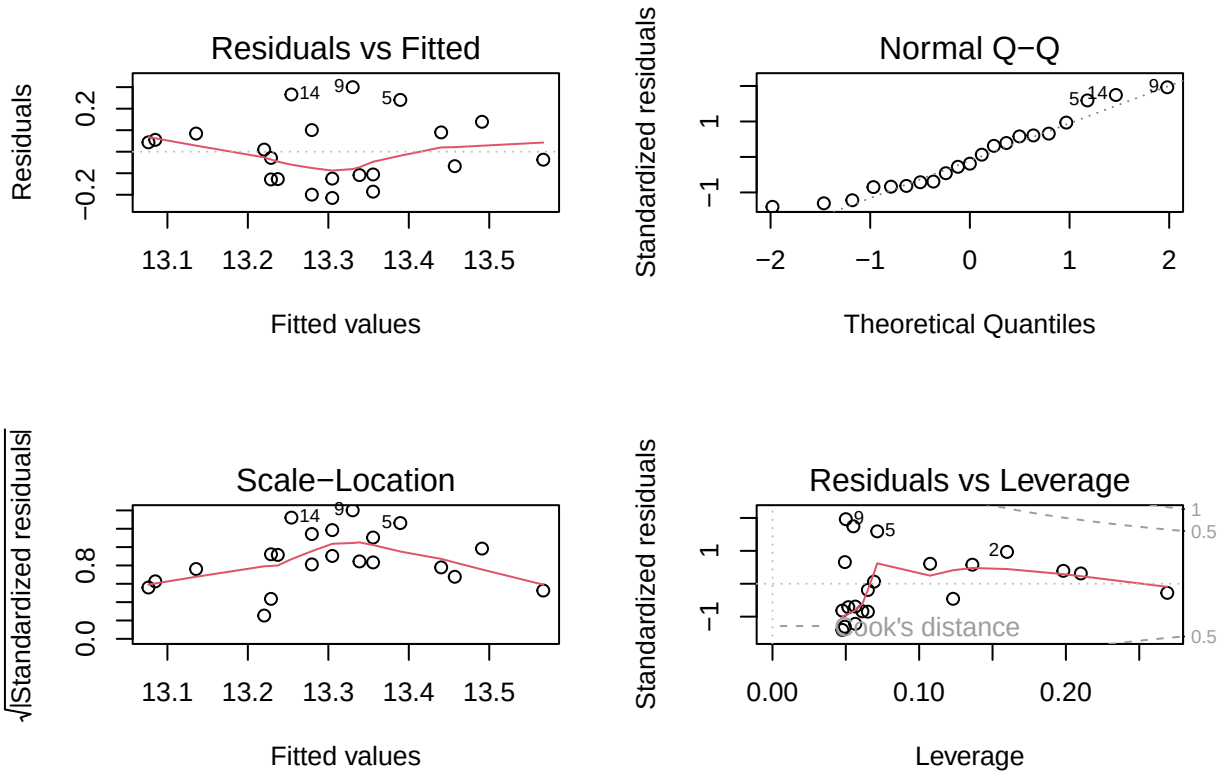


Model parameters estimates

```
##
## Call:
## lm(formula = time ~ windspeed, data = COLLIN)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21487 -0.12487 -0.02873  0.08976  0.29975
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.32179    0.03460 385.043  < 2e-16 ***
## windspeed   -0.08460    0.02361  -3.584  0.00198 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1567 on 19 degrees of freedom
## Multiple R-squared:  0.4033, Adjusted R-squared:  0.3719
## F-statistic: 12.84 on 1 and 19 DF,  p-value: 0.00198
```

The expected 110 metres hurdles run result in no wind (windspeed = 0 m/s) is 13.32 seconds. 1 m/s increase in wind speed results in almost 8.5 hundreths decrease in 110 metres hurdles run result.

Linear model assumptions check



The points on Residuals vs Fitted plot are randomly distributed around the value of 0. Smoother is roughly on the x-axis.

Normal Q-Q plot shows points approximately on the dashed line.

The model fits the data reasonably well.

Estimators explained including confidence intervals

```
##                   2.5 %      97.5 %
## (Intercept) 13.2493772 13.39420677
## windspeed   -0.1340109 -0.03519423
```

With 95% confidence, we expect the time/speed of the 110 meters in no wind to fall between 13.25 and 13.39 seconds.

If the wind speed increases by 1 m/s, we expect with 95% confidence that the time/speed of the 110 meter run will decrease between 3.5 and 13.4 hundredths of a second.

Coefficient of determination

It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.

```
## [1] 0.4033494
```

The coefficient of determination R2 has a value of 1 if all the variability for y is explained by the regression model.

In our case, the value of R2 is 0.4033; so 40% of the variability in y is explained by the regression model.

The two null hypotheses being tested and results explained.

```
CollinSum$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 13.32179201 0.03459815 385.043440 1.904681e-38
## windspeed   -0.08460258 0.02360620  -3.583913 1.979578e-03
```

The first null hypothesis assumes y is independent of x and the graph would give us a horizontal line. In our case, we reject the null hypothesis with 95% confidence. the p-value is smaller than 0.5 and we accept the conclusion that y depends on x.

In this case, we can also observe the null hypothesis for beta0 being 0. The value when all predictors of the model are equal to 0. We also reject this null hypothesis and assume with 95% confidence that beta0 is characteristic.

In no wind, the time/result of the 110 meter run would be 13.32 m/s.

Predicted values for chosen wind speeds

```
# izbrane vrednosti napovedne spremenljivke
cas_teka.napovedi<-data.frame(windspeed=c(-1,0,1,4))

# povprečne napovedi
povp.napovedi.Collin<-predict(CollinModel, cas_teka.napovedi, interval="confidence")
average_predict = data.frame(cbind(cas_teka.napovedi,povp.napovedi.Collin ))
average_predict
```

```
##   windspeed      fit      lwr      upr
## 1        -1 13.40639 13.31270 13.50008
## 2         0 13.32179 13.24938 13.39421
## 3         1 13.23719 13.15600 13.31838
## 4         4 12.98338 12.78355 13.18321
```

```
# posamične napovedi
pos.napovedi.Collin<-predict(CollinModel, cas_teka.napovedi, interval="prediction")
individual_predict = data.frame(cbind(cas_teka.napovedi,pos.napovedi.Collin ))
individual_predict
```

```
##   windspeed      fit      lwr      upr
## 1        -1 13.40639 13.06532 13.74747
## 2         0 13.32179 12.98594 13.65765
## 3         1 13.23719 12.89933 13.57504
## 4         4 12.98338 12.59934 13.36742
```

As you can see, individual predicts have broader confidence intervals.

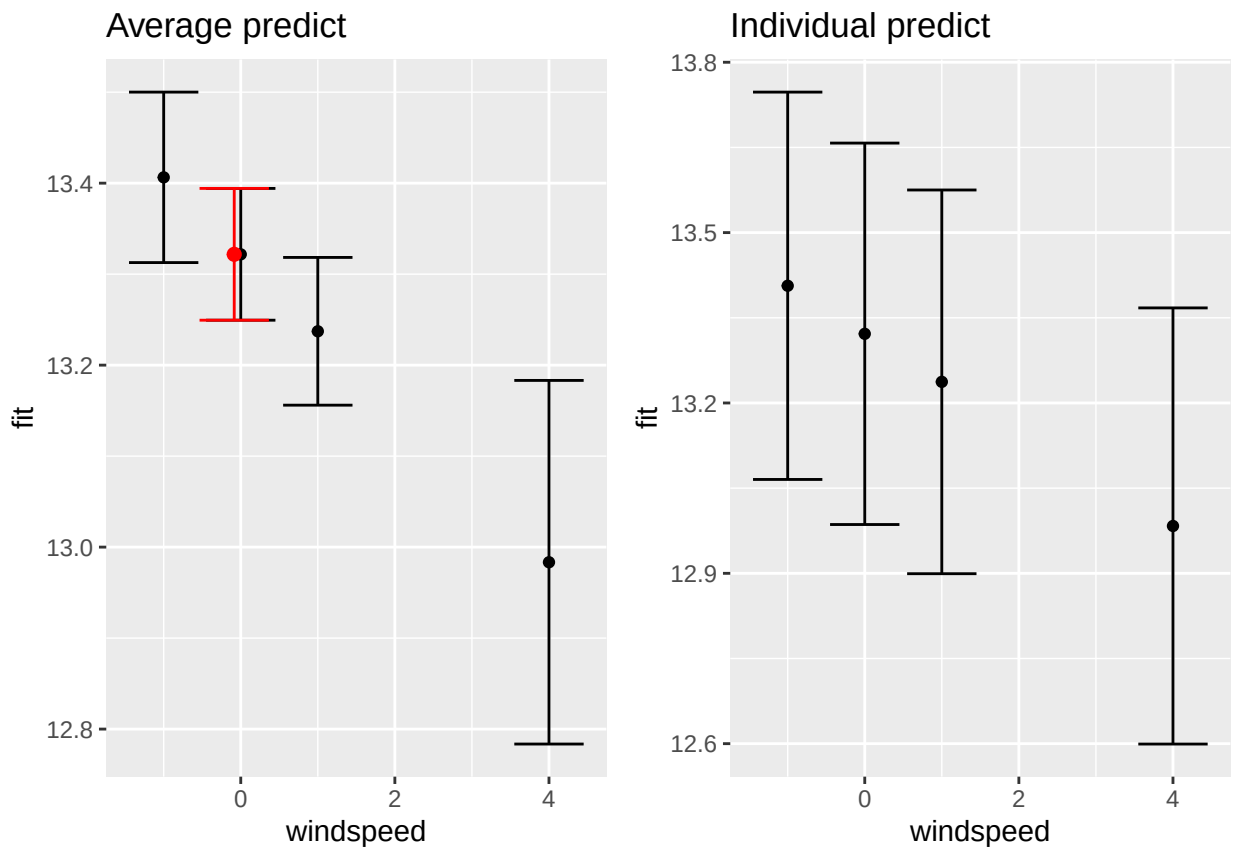
```
Confint = confint(CollinModel)
Confint = data.frame(Confint)
Confint1 = data.frame(windspeed=c(CollinModel$coefficients[[2]]), lwr = c(Confint$X2.5..[[1]]), fit = C

colcol = CollinModel$coefficients[[2]]

plot1 = ggplot(average_predict, aes(x=windspeed, y = fit))+
  geom_point()+
  geom_errorbar(aes(ymin = `lwr`, ymax = `upr`))+
  geom_errorbar(data = Confint1, aes(ymin = `lwr`, ymax = `upr`),color = "red")+
  geom_point(data = Confint1, aes(windspeed, fit, group=1),color = "red", size= 2)+
  ggtitle("Average predict")

plot2 = ggplot(individual_predict, aes(x=windspeed, y = fit))+
  geom_point()+
  geom_errorbar(aes(ymin = `lwr`, ymax = `upr`))+
  ggtitle("Individual predict")

grid.arrange(plot1, plot2, nrow=1, ncol=2)
```



Conclusion

Simple linear regression is used to model the relationship between two continuous variables. The objective is to predict the value of an output variable (or response - in our case run time/result) based on the value of an input (or predictor - in our case wind speed) variable.