Regression

Grusha Prasad

2024-10-28

One predictor

Generate some data

```
x = c(1:100)
y1 = x*0.5 + rnorm(100, 0, 5)
y2 = x*-2 + rnorm(100, 0, 5)
y3 = rnorm(100, 0, 5)

dat= data.frame(x=x,
                y1=y1,
                y2=y2,
                y3=y3)
```

Fit regression model

Here is code that fits regression models for each of the three y values. What do you notice about the coefficient values and the p-values? Does this make sense?

```
fit1 = lm(y1 ~ x, data = dat)
summary(fit1)
##
## Call:
## lm(formula = y1 ~ x, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.4694  -2.7619   0.8994   3.5885  13.3229
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.37474    1.02109  -0.367    0.714
## x            0.50367    0.01755  28.692   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.067 on 98 degrees of freedom
## Multiple R-squared:  0.8936, Adjusted R-squared:  0.8925
## F-statistic: 823.2 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
fit2 = lm(y2 ~ x, data = dat)
summary(fit2)
```

```
##
## Call:
## lm(formula = y2 ~ x, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.405  -3.448  -0.311   3.165  11.247
##
## Coefficients:
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  0.16264    1.05067    0.155    0.877
## x           -2.00572    0.01806 -111.042   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.214 on 98 degrees of freedom
## Multiple R-squared:  0.9921, Adjusted R-squared:  0.992
## F-statistic: 1.233e+04 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
fit3 = lm(y3 ~ x, data = dat)
summary(fit3)
##
## Call:
## lm(formula = y3 ~ x, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.785  -2.881   0.510   3.051  11.380
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.827080   1.050551  -0.787    0.433
## x            0.004936   0.018061   0.273    0.785
##
## Residual standard error: 5.213 on 98 degrees of freedom
## Multiple R-squared:  0.0007616,  Adjusted R-squared:  -0.009435
## F-statistic: 0.0747 on 1 and 98 DF,  p-value: 0.7852
```

Two predictors

Here is code that will generate y values that are a combination of two predictors, x1 and x2.

```
x1 = c(1:100)
x2 = rep(c(1, 0), times=50)

y1 = x1*0.5 + x2*-20 + rnorm(100, 0, 5)
y2 = x*-2 + x2*0.01 + rnorm(100, 0, 10)

dat2= data.frame(x=x,
                y1=y1,
                y2=y2)
```

Fit regression model

Here is code that fits regression models with two predictors. What do you notice about the coefficient values and the p-values? Specifically consider p-value of x2 in fit5. Does this make sense?

```
fit4 = lm(y1 ~ x1 + x2, data=dat2)
summary(fit4)
##
## Call:
## lm(formula = y1 ~ x1 + x2, data = dat2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.4429  -3.2209   0.0508   4.0914  11.0661
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.8661     1.1300  -2.536   0.0128 *
## x1            0.5284     0.0173  30.543   <2e-16 ***
## x2          -18.6310     0.9988 -18.654   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.993 on 97 degrees of freedom
## Multiple R-squared:  0.9306, Adjusted R-squared:  0.9292
## F-statistic: 650.5 on 2 and 97 DF,  p-value: < 2.2e-16
```

```
fit5 = lm(y2 ~ x1 + x2, data=dat2)
summary(fit5)
##
## Call:
## lm(formula = y2 ~ x1 + x2, data = dat2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.9940  -5.6854  -0.8053   6.0003  20.1652
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.09651    2.04945   1.023    0.309
## x1          -2.01189    0.03137 -64.125   <2e-16 ***
## x2           0.27550    1.81131   0.152    0.879
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.055 on 97 degrees of freedom
## Multiple R-squared:  0.977,  Adjusted R-squared:  0.9765
## F-statistic:  2057 on 2 and 97 DF,  p-value: < 2.2e-16
```