# R projects 7 and 8

Call:

lm(formula = emass ~ dur, data = echidna)

Residuals:

 Min 1Q Median 3Q Max

-0.74428 -0.32936 -0.01557 0.29868 0.78492

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.968240 0.302352 16.432 6.54e-16 \*\*\*

dur -0.010730 0.002413 -4.447 0.000125 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4376 on 28 degrees of freedom

Multiple R-squared: 0.414, Adjusted R-squared: 0.3931

F-statistic: 19.78 on 1 and 28 DF, p-value: 0.0001254

The regression line has equation: y= 4.968240-0.010730x. Thus, we expect the mass of an echidna that hibernates for 150 days to be about 4.968240-0.010730(150)=3.359 kg.

Checking to see if the residuals are plausibly normally distributed with mean 0 and a common variance:

* There’s no obvious pattern in the Residuals vs Fitted plot.
* The Normal Q-Q plot shows some consistent departure from the line at the lower tail. This is fairly minor, though it would prompt me to check some of the more extreme measurements: perhaps there was measurement error or some circumstance (disease, flooding of burrow, etc.) that might lead me to exclude certain observations from the model.

The model seems reasonable and so we have highly significant evidence (p-value 0.000125) for a link between emergence mass and hibernation duration. Out estimate of R-squared is 0.414, meaning that 41.4% of the variation in emergence mass is due to variation in hibernation duration. The correlation is negative.

Call:

lm(formula = dur ~ emass, data = echidna)

Residuals:

 Min 1Q Median 3Q Max

-40.705 -19.564 -2.493 17.969 51.492

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 262.512 32.207 8.151 7.13e-09 \*\*\*

emass -38.582 8.675 -4.447 0.000125 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.24 on 28 degrees of freedom

Multiple R-squared: 0.414, Adjusted R-squared: 0.3931

F-statistic: 19.78 on 1 and 28 DF, p-value: 0.0001254

The regression line has equation: y= 262.512-38.582x, thus we expect the hibernation duration of an echidna with an emergence mass of 4 kg to have been about 262.512-38.582(4)=108 days.

Checking to see if the residuals are plausibly normally distributed with mean 0 and a common variance: these plots are essentially the same as for the linear model in the other direction. Note that we have exactly the same p-value and R-squared. The correlation is still negative.

Call:

lm(formula = e4 ~ e1)

Residuals:

 Min 1Q Median 3Q Max

-33.674 -5.261 1.626 7.764 22.590

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 21.3822 8.9714 2.383 0.0209 \*

e1 0.7125 0.1119 6.369 5.4e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.89 on 51 degrees of freedom

 (7 observations deleted due to missingness)

Multiple R-squared: 0.443, Adjusted R-squared: 0.4321

F-statistic: 40.56 on 1 and 51 DF, p-value: 5.396e-08

Based on the R-squared of 0.443 and the generally good fit of the regression line to the data, it would not be unreasonable to drop exam 4. This would not be likely to change grades on average, however individual students might have very different grades. These individuals are the points in the plot far from the regression line.

Analysis of the residuals shows some cause for concern: variation seems to be smaller for higher scores (seen the Fitted vs Residuals plot) and there’s a significant drop below the line in the Normal-QQ plot. However, the model seems generally okay and I think we have strong evidence for a connection between the exam scores.



Call:

lm(formula = LE ~ H, data = cia)

Residuals:

 Min 1Q Median 3Q Max

-19.404 -5.809 2.102 6.294 13.181

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 64.8541 1.7473 37.116 < 2e-16 \*\*\*

H 0.7123 0.2397 2.971 0.00353 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.042 on 132 degrees of freedom

Multiple R-squared: 0.06267, Adjusted R-squared: 0.05557

F-statistic: 8.826 on 1 and 132 DF, p-value: 0.003529

We see a weak positive correlation between health-care expenditures and life expectancy (R-squared is just 0.06267). The Normal Q-Q plot of the residuals shows some series wiggles, so we should be cautious with the p-value of 0.00353. However, I think it’s still safe to conclude that there is some connection between health-care expenditures and life expectancy.

Call:

lm(formula = LE ~ CU, data = cia)

Residuals:

 Min 1Q Median 3Q Max

-21.8831 -2.9337 0.3237 3.7005 17.7815

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 57.35633 1.32539 43.27 <2e-16 \*\*\*

CU 0.25154 0.02485 10.12 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.233 on 132 degrees of freedom

Multiple R-squared: 0.4371, Adjusted R-squared: 0.4328

F-statistic: 102.5 on 1 and 132 DF, p-value: < 2.2e-16

This time we see a much stronger positive correlation (R-squared is 0.4371). Again the Normal Q-Q plot of residuals shows some cause for concern. However, this data is clearly showing a much stronger connection between contraceptive use and life expectancy. Based on this, it seems that increasing the use of contraception is a more reliable way of increasing life expectancy than is spending more on health care. This may be because more contraceptive use means that people can make more careful choices about when they become pregnant and thus maternal and childhood mortality decreases. It may also be that there is a confounding variable: some other thing that is related to both. Perhaps countries with more gender equality have both higher rates of contraceptive use and higher life expectancy because women get better (more equal) health care.

Call:

lm(formula = m ~ f, data = opah)

Residuals:

 Min 1Q Median 3Q Max

-11.4718 -2.6969 -0.6542 3.0093 12.2469

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) -73.04427 4.79306 -15.24 <2e-16 \*\*\*

f 1.16353 0.04792 24.28 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.421 on 89 degrees of freedom

Multiple R-squared: 0.8688, Adjusted R-squared: 0.8674

F-statistic: 589.6 on 1 and 89 DF, p-value: < 2.2e-16

Although this seems like a pretty good fit, the plot of Residuals vs Fitted shows signs of a pattern and the Normal Q-Q plot of residuals is pretty wiggly. It makes sense to look at logarithms because mass should be roughly proportional to volume, which in turn should be roughly proportional to the cube of the fork length:

mass = C\*(F^3) for some constant C. Thus log(mass) = logC + 3logF. This is a linear relationship between the logarithms of the mass and the fork length.

Call:

lm(formula = log(m) ~ log(f), data = opah)

Residuals:

 Min 1Q Median 3Q Max

-0.272676 -0.069112 -0.007057 0.051664 0.225111

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) -9.26697 0.44877 -20.65 <2e-16 \*\*\*

log(f) 2.82466 0.09762 28.93 <2e-16 \*\*\*

---

Residual standard error: 0.09255 on 89 degrees of freedom

Multiple R-squared: 0.9039, Adjusted R-squared: 0.9028

F-statistic: 837.2 on 1 and 89 DF, p-value: < 2.2e-16