17.7: Regression model fit

Introduction

In Chapter 17.5 and 17.6 we introduced the example of tadpoles body size and oxygen consumption. We ran a simple linear regression, with the following output from R

```
RegModel.1 <- lm(VO2~Body.mass, data=example.Tadpole)

summary(RegModel.1)

Call:
lm(formula = VO2 ~ Body.mass, data = example.Tadpole)

Residuals:
    Min      1Q    Median       3Q       Max
-202.26 -126.35     30.20    94.01    222.55

Coefficients:
                Estimate     Std. Error    t value    Pr(>|t|)
(Intercept)      -583.05         163.97     -3.556     0.00451 **
Body.mass         444.95          65.89      6.753   0.0000314 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 145.3 on 11 degrees of freedom
Multiple R-squared: 0.8057, Adjusted R-squared: 0.788
F-statistic: 45.61 on 1 and 11 DF, p-value: 0.00003144
```

You should be able to pick out the estimates of slope and intercept from the table (intercept was -583 and slope was 445). Additionally, as part of your interpretation of the model, you should be able to report how much variation in VO2 was explained by tadpole body mass (**coefficient of determination**, R^2, was 0.81, which means about 81% of variation in oxygen consumption by tadpoles is explained by knowing the body mass of the tadpole.

What's left to do? We need to evaluate how well our model fits the data, i.e., we evaluate regression model fit. This we can do by evaluating the error components relative to the portion of the model that explains the data. Additionally, we can perform a number of diagnostics of the model relative to the assumptions we made to perform linear regression. These diagnostics form the subject of Chapter 17.8. Here, we ask how well does the model

$$\dot{V}O_2 = b_0 + b_1(Bodymass)$$

fit the data?

Model fit statistics

The second part of fitting a model is to report how well the model fits the data. The next sections apply to this aspect of model fitting. The first area to focus on is the magnitude of the residuals: the greater the spread of residuals, the less well a fitted line explains the data.

In addition to the output from `lm()` function, which focuses on the coefficients, we typically generate the ANOVA table also.

```
Anova(RegModel.1, type="II")
Anova Table (Type II tests)
```

```
Response: VO2
               Sum Sq   Df    F value       Pr(>F)
Body.mass      962870    1     45.605   0.00003144 ***
Residuals      232245   11
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Standard error of regression

S, the **Residual Standard Error** (aka **Standard error of regression**), is an overall measure to indicate the accuracy of the fitted line: it tells us how good the regression is in predicting the dependence of response variable on the independent variable. A large value for S indicates a poor fit. One equation for **S** is given by

$$\mathbf{S} = \sqrt{\frac{SS_{residual}}{n-2}}$$

In the above example, $\mathbf{S} = 145.3$ (underlined, bold in regression output above). We can see how if $SS_{residual}$ is large, **S** will be large indicating poor fit of the linear model to the data. However, by itself **S** is not of much value as a diagnostic as it is difficult to know what to make of 145.3, for example. Is this a large value for **S** ? Is it small? We don't have any context to judge **S**, so additional diagnostics have been developed.

Coefficient of determination

R^2, the **coefficient of determination**, is also used to describe model fit. R^2, the square of the simple product moment correlation r, can take on values between 0 and 1 (0% to 100%). A good model fit has a high R^2 value. In our example above, $R^2 = 0.8057$ or 80.57%. One equation for R^2 is given by

$$R^2 = \frac{SS_{regression}}{SS_{total}}$$

A value of R^2 close to 1 means that the regression "explains" nearly all of the variation in the response variable, and would indicate the model is a good fit to the data. Note that the coefficient of determination, R^2, is the squared value of r, the product moment correlation.

Adjusted R-squared

Before moving on we need to remark on the difference between R^2 and adjusted R^2. For Simple Linear Regression there is but one predictor variable, X; for multiple regression there can be many additional predictor variables. Without some correction, R^2 will increase with each additional predictor variables. This doesn't mean the model is more useful, however, and in particular, one cannot compare R^2 between models with different numbers of predictors. Therefore, an adjustment is used so that the coefficient of determination remains a useful way to assess how reliable a model is and to permit comparisons of models. Thus, we have the Adjusted $\bar{R}^2$, which is calculated as

$$\bar{R}^2 = 1 - \frac{SS_{residual}}{SS_{total}} \cdot \frac{DF_{total}}{DF_{residual}}$$

In our example above, Adjusted $R^2 = 0.3806$ or 38.06%.

Which should you report? Adjusted R^2, because it is independent of the number of parameters in the model.

Both $\bar{R}^2$ and **S** are useful for regression diagnostics, a topic which we will discuss next (Chapter 17.8).

Questions

1. True or False. The simple linear regression is called a "best fit" line because it maximizes the squared deviations for the difference between observed and predicted Y values.
2. True or False. Residuals in regression analysis are best viewed as errors committed by the researcher. If the experiment was designed better, or if the instrument was properly calibrated, then residuals would be reduced. Explain your choice.

3. The USA is finishing the 2020 census as I write this note. As you know, the census is used to reapportion Congress and also to determine the number of electoral college votes. In honor of the election for US President that's just days away, in the next series of questions in this Chapter and subsequent sections of Chapter 17 and 18, I'll ask you to conduct a regression analysis on the electoral college. For starters, make the regression of Electoral votes on the 2010 census population. (Ignore for now the other columns, just focus on POP_2019 and Electoral.) Report the
    - regression coefficients (slope, intercept)
    - percent of the variation in electoral college votes explained by the regression $\left(R^2\right)$.
4. Make a scatterplot and add the regression line to the plot

Data set

State	Region	Division	POP_2010	POP_2019	Electoral
Alabama	South	East South Central	4779736	4903185	9
Alaska	West	Pacific	710231	731545	3
Arizona	West	Mountain	6392017	7278717	11
Arkansas	South	West South Central	2915918	3017804	6
California	West	Pacific	37253956	39512223	55
Colorado	West	Mountain	5029196	5758736	9
Connecticut	Northeast	New England	3574097	3565287	7
Delaware	South	South Atlantic	897934	982895	3
District of Columbia	South	South Atlantic	601723	705749	3
Florida	South	South Atlantic	18801310	21477737	29
Georgia	South	South Atlantic	9687653	10617423	16
Hawaii	West	Pacific	1360301	1415872	4
Idaho	West	Mountain	1567582	1787065	4
Illinois	Midwest	East North Central	12830632	12671821	20
Indiana	Midwest	East North Central	6483802	6732219	11
Iowa	Midwest	West North Central	3046355	3155070	6
Kansas	Midwest	West North Central	2853118	2913314	6
Kentucky	South	East South Central	4339367	4467673	8
Louisiana	South	West South Central	4533372	4648794	8
Maine	Northeast	New England	1328361	1344212	4
Maryland	South	South Atlantic	5773552	6045680	10
Massachusetts	Northeast	New England	6547629	6892503	11
Michigan	Midwest	East North Central	9883640	9883635	16
Minnesota	Midwest	West North Central	5303925	5639632	10
Mississippi	South	East South Central	2967297	2976149	6
Missouri	Midwest	West North Central	5988927	6137428	10
Montana	West	Mountain	989415	1068778	3

State	Region	Division	POP_2010	POP_2019	Electoral
Nebraska	Midwest	West North Central	1826341	1934408	5
Nevada	West	Mountain	2700551	3080156	6
New Hampshire	Northeast	New England	1316470	1359711	4
New Jersey	Northeast	Mid-Atlantic	8791894	8882190	14
New Mexico	West	Mountain	2059179	2096829	5
New York	Northeast	Mid-Atlantic	19378102	19453561	29
North Carolina	South	South Atlantic	9535483	10488084	15
North Dakota	Midwest	West North Central	672591	762062	3
Ohio	Midwest	East North Central	11536504	11689100	18
Oklahoma	South	West South Central	3751351	3956971	7
Oregon	West	Pacific	3831074	4217737	7
Pennsylvania	Northeast	Mid-Atlantic	12702379	12801989	20
Rhode Island	Northeast	New-England	1052567	1059361	4
South Carolina	South	South-Atlantic	4625364	5148714	9
South Dakota	Midwest	West-North-Central	814180	884659	3
Tennessee	South	East-South-Central	6346105	6829174	11
Texas	South	West-South-Central	25145561	28995881	38
Utah	West	Mountain	2763885	3205958	6
Vermont	Northeast	New-England	625741	623989	3
Virginia	South	South-Atlantic	8001024	8535519	13
Washington	West	Pacific	6724540	7614893	12
West Virginia	South	South-Atlantic	1852994	1792147	5
Wisconsin	Midwest	East-North-Central	5686986	5822434	10
Wyoming	West	Mountain	563626	578759	3