Regression Analysis - STAT 482 - HW #2

Alex Towell (atowell@siue.edu)

Refer to the data from Exercise 1.27

A person's muscle mass is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 15 women for each 10 year age group, beginning with 40 and ending with age 79. The input variable x is age (in years), and the response variable y is muscle mass (in muscle mass units).

Part (1)

Compute the t-statistic and p-value for testing $H_0 : \beta_1 = 0$.

We assume a random sample $\{\mathcal{D}_i\}$ where $\mathcal{D}_i = (X_i, Y_i)$ is generated by the linear model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$. We are not particularly interested in the distribution of X_i and instead seek to use it as an explanatory variable (input) for response variable Y_i (output).

Thus, we are interested in estimating the parameters of the conditional distribution

$$Y_i | X_i = x_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2).$$

We fit the given data $\{d_i\}$ where $d_i = (x_i, y_i)$ to the linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

using the following R code:

```
mass.data = read.table('CH01PR27.txt')
colnames(mass.data)=c("mass","age")
mass.mod = lm(mass ~ age, data=mass.data)
```

We compute the t-statistic $t^* = \frac{b_1}{\text{SE}(b_1)} \sim t(n-2)$ and the p-value $p = \Pr\{t(n-2) > |t_*|\}$ with the following R code:

```
beta1.stats = summary(mass.mod)$coefficients[2,]
beta1.stats
```

```
##      Estimate    Std. Error       t value      Pr(>|t|)
## -1.189996e+00  9.019725e-02 -1.319326e+01  4.123987e-19
```

We see that $t^* = -13.193$ which has a p-value less than 0.001.

Part (2)

Provide an interpretation of the above as a test for model compatibility.

We think of the the null hypothesis as a simplification to the model, which we *a priori* prefer to the more complicated model (there is a probabilistic justification for this bias),

$$H_0 : \beta_1 = 0 \qquad \text{(age has no effect on muscle mass)}.$$

We obtained a p-value of approximately .000, and so the data is not compatible with the no effect model; we accept the model which includes age as a predictor for muscle mass.

Alterantive formulation

Can we say the null hypothesis $H_0 : \beta_1 = 0$ is equivalent to

$$H_0 : \text{no association between age and muscle mass}$$

since if $\beta_1 = 0$ then

$$Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\beta_0, \sigma^2)?$$

However, I am concerned that I am going beyond the model since our model is given by the conditional distribution of $Y_i|X_i$. So, we may say that x_i has no effect on Y_i, but in general y_i may effect X_i. Clearly, in this case, we have scientific insight that one's muscle mass does not effect one's age.

Also, I'm struggling with the wording *compatibility*. I find myself wanting to say that any outcome that has non-zero probability under the model is compatible with the model. Is it simply too subtle to say, "Assuming the null hypothesis is true, then with probability p the observed statistic or a more extreme value will have been observed." This may in fact be a bit much, especially for the less mathematically inclined, whom we wish to inform with our analysis. So, perhaps we are better off using phrasing like "fail to reject the null hypothesis" or "the data is not compatible with the null model." I must confess that I sort of prefer "fail to reject the null hypothesis," since it seems to be taking the stance that either model may be correct, but we lack sufficient data to reject the null model, which is some model we prefer for some reason, e.g., the null model is often the simplest model in a family of models.

I'm still unsatisfied with my thoughts on this subject. On a related note, I enjoyed your discussion last Spring on the Bayesian perspective. When we combine this perspective with, say, a universal prior, perhaps we are getting closer to a sort of formalism for the scientific method (where the universal prior is a sort of formalism for Occam's razor).

Part (3)

Provide an interpretation of the above as a test for sign / direction of the effect.

Since both $t^* < 0$ and $b_1 < 0$, the data supports the hypothesis of a negative association between age and muscle mass, i.e., as age increases muscle mass is expected to decrease.

Part (4)

Compute an interval estimate for β_1.

We compute a 95% confidence interval for β_1 with:

```
alpha = 0.05
beta1.ci = round(confint(mass.mod, level=1-alpha)[2,],digits=3)
beta1.ci
```

```
##  2.5 % 97.5 %
## -1.371 -1.009
```

We see that β_1 has a 95% confidence interval given by

$$[-1.371, -1.009].$$

Part (5)

Explain how an interval estimate provides additional information beyond the test result.

We can say more about the size of the effect and a measure of the estimation accuracy.

We make two observations about the CI for β_1:

1. All the values in the CI are distant from 0, and thus we can say there is strong evidence that age has a large effect.
2. The CI represents a relatively small range of values, and thus we can say the model is accurate, i.e., the data has a significant amount of information about β_1.

Part (6)

Compute an interval estimate for the centered model parameter β_0^*.

β_0^* represents the mean muscle mass at the mean age in the sample,

$$E(Y|x = \bar{x}) = \beta_0^*,$$

and thus

$$b_0^* = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

To compute the confidence interval for b_0^*, I refit the model using the following R code:

```
x = mass.data$age
y = mass.data$mass
x.bar = mean(x)
y.bar = mean(y)
x.star = x - x.bar
star.mod = lm(y ~ x.star, data=mass.data)
beta0.star.ci = round(confint(star.mod)[1,],digits=3)
beta0.star.ci
```

```
##  2.5 % 97.5 %
## 82.855 87.079
```

We see that β_0^* has a 95% confidence interval given by

$$[82.855, 87.079].$$

Part (7)

Provide an interpretation for your result, stated in the context of the problem.

We estimate that the mean muscle mass for all women at the center value of age ($\bar{x} = 59.983$) is between 82.855 and 87.079.