1 General Notes

There is a model on mother earnings ranges divided into 4 scales (see Gelman page 67) once the model has been fit, the following coefficients have been estimated:

$$s = 82.0 + 3.8 \cdot w_2 + 11.5 \cdot w_3 + 5.2 \cdot w_4 \tag{1}$$

Observe that w_1 is missing from the equation, and its predicted score value is the intercept (and other values being equal to zero). There is a nice example in Gelman page 53 on how different units affect coefficient interpretation. Compare the following

$$earnings = -61000 + 51 \cdot height[inmillimetres] + \epsilon \tag{2}$$

And

$$earnings = -61000 + 81000000 \cdot height[inmiles] + \epsilon \tag{3}$$

It has been given that standard deviation of height is equal to 3.8 inches, which is 97 millimetres or 0.000061 miles. Observe that we obtain the same expected difference in earnings for the matching units

$$51 \cdot 97 = 81000000 \cdot 0.000061 = 4900 \tag{4}$$

2 Scaling

Suppose there is this model

$$\log_e arn = height + male \tag{5}$$

Once we fit a linear regression model we have received the following coefficients:

$$\log_e arn = 8.153 + 0.021 height + 0.423 male \tag{6}$$

Also,

$$\exp(0.021) = 1.02 \tag{7}$$

Therefore, for two people of the same gender one inch of height contributes towards 2% of salary increase. Say that the standard errors are as follows:

intercept	height	male
0.603	0.009	0.072

The regression model has a residual standard deviation of 0.88, implying that approximately 68% of log earnings will be within 0.88 of the predicted value. As a result, a 70-inch tall person will have earnings equal to $8.153+0.021 \cdot 70 = 9.623$. Knowing that predictive standard deviation is 0.88, there is 68 chance the person earning are within $9.623 + / - 0.88 = [8.74, 10.50] = [\exp(8.74), \exp(10.50)] = [6000, 36000]$. The R^2 value for this model was pretty low therefore it is expected to observe such a wide salary range.

2.1 General approach to standarization

On another note, we should always standarize the input data to make the interpretation easier.As a result, when analysing the output coefficients, the requirements for other variables being zero is equal to say they are at their averages. There is an example on page 96 that talks about this scenario.

$$s = \beta_0 +_1 \times d +_2 \times a +_3 \times da, \tag{8}$$

where s is the probability of switching between wells, d is the distance to the closest well, a is the arsenic level and da is the interaction term between distance and arsenic level.

The model came out to be

$$s = 0.35 + -0.88 \times d + 0.47 \times a - 0.18 \times da, \tag{9}$$

Therefore, the probability of switching when the distance and arsenic level are at their average is equal to $logit^{-1}(0.35) = 0.59$.

3 Interaction

Consider the following model

$$\log(earnings)\ (height) \tag{10}$$

Once we have fit the equation coefficients given data we obtain the following:

1. intercept - 5.74
2. height - 0.06

The difference of 0.06 in height corresponds to $\exp(0.06) = 1.062$. Therefore, a unite change of β_1 corresponds to a 6% increase in the y value. The opposite holds for negative values (i.e there is 6% decrease in the outcome value for a drop of one unit in β_1). Observe that It would be interesting to see if gender distinction contributes to the 6% increase.

Suppose there is this equation

$$\log(e) = 8.4 + 0.017 \cdot h - 0.079 \cdot m + 0.007 \cdot h \cdot m \tag{11}$$

where e corresponds to earnings, h height, m to male and $h \cdot m$ it an interaction term between height and male. Now, observe that male intercept does not really have a proper interpretation as it expects the height being equal to 0. However, the interaction term corresponds to the " difference in the slopes for log earnings between men and women". Therefore, in interaction term gives us 0.7% increase for men for an inch increase in their height. Moreover, for men and inch increase in hight predicts $1.7\% + 0.7\% = 2.4\%$

4 Logistic Regression

Gelman notes on page 81 that "As with linear regression, the intercept can only be interpreted assuming zero values for the other predictors." Observe that this enforces the analyst to look at another point of reference. In other words, if there is equation with only one predictor x, we could analyse the predicted value at $\bar{x}$, or any other meaningful point. By meaningful, we are saying values that are observed in the data. In the example provided in the book the equation is given as

$$logit^{-1} = -1.40 + 0.33x \tag{12}$$

What we can do is pick two consecutive units of x and then report the percentage change, given the unit difference

It is assumed on page 83 that coefficient estimates within two standard errors from the estimated coefficient value are consistent with data. However, a statistically significant coefficient is required to be at least 2 standard errors **away from zero**.

This is also reported on pages 90-91 when there is a following model presented:

$$s = -0.9 \cdot d + 0.46 \cdot a, \tag{13}$$

where s is probability of switching between two wells, d is the distance to the nearest well in meters (divided by 100), a is the arsenic level of the currently used well. Intercept has been estimated as very close to 0. With the quick application of the dive by 4 rule, we could see that for the same arsenic level, $100m$ corresponds to -0.22 probability of switching. Similarly, one unit difference on the arsenic scale corresponds to 11% to probability of switching. What this means is that the closer the safe well is - the easier it is for household to switch. In addition, the higher the arsenic level, the more probable it is to switch.

It is very important to note that we cannot claim that distance is more of an influential factor than arsenic level.What we should do instead is to investigate how one standard deviation change in data affects the probability of switching. As a result, we will be able to report on differences in "common units". Standard deviations are equal to $\sigma_d = 0.38$, $\sigma_a = 1.1$, which yields $-0.9 \cdot 0.38 = -0.34$, and $0.46 * 1.10 = 0.51$. Now, we can apply the divide by 4 rule on the estimated data. Thus, a σ_d change in distance corresponds to -8% (negative) difference to the probability of switching, while a σ_a change in arsenic level gives 13% in the probability of switching.

4.0.1 UCLA datasets

There is a good introduction to logistic regression on UCLA website. There is a dataset that contains test results split between two genders. This is the gender distribution:

1. female - 109

2. male - 91

The first analysis is to run a logistic regression to check how many students finish their class with honorous distinction.

	H	N
F	32	77
M	17	74

We would like to run a logistic model to regress gender onto class type:

$$h = \beta_0 + \beta_1 g \tag{14}$$

where h represents the class that students attended and g is gender. The equation is estimated as

$$h = -1.4709 + 0.5928g \tag{15}$$

Observe that $\exp(-1.4709) = 0.2297$. The odds of being in honorous class for males is $\frac{17}{74} = 0.2297$, which is the same as the exponent of the intercept. Therefore, the estimated intercept is equal to ratio of the reference group (i.e. male) belonging to the honorous class.

Recall, that the ratio of female students attending a honorous class is $\frac{32}{77}$. As a result, the ratio of the odds of females to males is

$$\frac{\frac{32}{77}}{\frac{17}{74}} = 1.8090 \tag{16}$$

, which is the same as the exponent of the coefficient of female $\exp(0.5928) = 1.8090$.

Subsequently, we could analyse a model in which math scored is regress on whether the students have attended a honorous class or not, as in:

$$h = \beta_0 + \beta_1 m, \tag{17}$$

where m is math score, and h a math score.

Obviously this works well for two binary options. Suppose there is $k-$level categorical variable denoting earnings and it regress on a candidate selection during national elections. Therefore, Gelman on pages 82-82 talks about mathematical approach given the estimated formula. These are the potential solutions

- "Divide by four" - it turns out that if we calculate β 4 we will get a pretty good estimate of what the percentage of change is given a unit change for a variable.
- Unit change - in this approach we plug in 2 consecutive units into the regression equation, and look at the percentage change.
- latent approach - defined as an observable set of pairs

$$(y_i, X_i\beta + \epsilon) \tag{18}$$

where $y \in \{0, 1\}$, and $y_i = 1$, when $X_i\beta + \epsilon > 0$ and $y_i = 0$ otherwise. Error terms are assumed to have logistic distribution. Also, error term follows Gaussian distribution with $\mu = 0$ and $\sigma = 1.6$

$$\epsilon N(0, \sigma^2) \tag{19}$$