Comments on homework, Correlation/Linear Regression

- Positive/negative and strong/moderate/weak correlation (might be easier to get right than high/low).
- Estimated (rather than population) parameters:
  - $Y_i = \alpha + \beta X_i + \epsilon_i$
  - $\hat{Y} = \hat{\alpha} + \hat{\beta} X$
  - $\hat{Y}_{totalpoints} = \hat{\alpha} + \hat{\beta} X_{pointslongjump}$
- R^2 - proportion of the total variance of the response (dependent) variable explained (more on it today).

Significance testing for the slope coefficient

- **Type** of data (**level** of measurement)
  Continuous response variable
- **Assumptions** about the data
  normally distributed error term ϵ with mean 0 and and a variance σ^2 which is the same for all units
- Statistical **hypotheses**:
  - Null hypothesis $H_0 : \beta = 0$
  - Alternative hypothesis $H_a : \beta \neq 0$
- **Test statistic** t-test
- **P-value** (using sampling distribution of the test statistic)
- Substantive **conclusion** (inference in the population)

Some useful formulas

- Test statistic:
  $t = \frac{\hat{\beta}}{\hat{se}(\hat{\beta})}$
- Degrees of freedom:
  df = n - (k + 1), where k is number of explanatory variables
- Confidence intervals (when n is large):
  $\hat{\beta} \pm z_{\alpha/2}\hat{se}(\hat{\beta})$
  $\hat{\beta} \pm 1.96\hat{se}(\hat{\beta})$
- Confidence intervals (general form):
  $\hat{\beta} \pm t_{\alpha/2}^{(n-(k+1))}\hat{se}(\hat{\beta})$

Linear association

between one's years of schooling and his/her parents

- $\hat{Y}_{educ} = \hat{\alpha} + \hat{\beta}_{paeduc} X_{paeduc}$
- $\hat{\alpha}$, intercept, constant
- $\hat{\beta}_{paeduc}$, slope, regression coefficient

Linear association

between one's years of schooling and his/her parents

- $\hat{\alpha} = 9.782$
- $\hat{\beta}_{paeduc} = 0.354$
- $t = \frac{0.354}{0.017} = 20.905$
- $p < 0.001$

Linear association

between one's years of schooling and his/her parents

- $\hat{Y}_{educ} = 9.782 + 0.354_{paeduc} X_{paeduc}$
- $\hat{\alpha} = 9.782$
- $\hat{\beta}_{paeduc} = 0.354$
- $t = \frac{0.354}{0.017} = 20.905$
- $p < 0.001$

Linear association

between one's years of schooling and his/her parents

The linear regression of highest year of school completed by the respondent on highest year of school completed by respondents father shows a linear association between the two attributes. For every extra year of school completed by a respondents father, we would expect the respondents own schooling to be 0.354 years higher. This is a statistically significant linear relationship, at any conventional level of significance (t=20.905, p¡0.001) we can confidently reject the null hypothesis that in the population there is no linear association between these attributes.