Introduction and Single Predictor Regression

Dr. J. Kyle Roberts

Southern Methodist University
Simmons School of Education and Human Development
Department of Teaching and Learning

Correlation

- A correlation is a symmetric, scale-invariant measure of the (linear) association between two random variables.
- The correlation is completely symmetric between the two variables. We do not assume that one is the predictor and the other is the response. In most cases we assume that both variables are being driven by an unobserved, "hidden" or "lurking" variable.
- In other words correlation between variables is an observed or empirical trait. It does not imply causation.

Pearson correlation

- The Pearson correlation

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} = \frac{COV_{xy}}{SD_x SD_y}$$

  is the most common measure of correlation.
- Both r and ρ are dimensionless and restricted to $[-1, 1]$.
- A correlation (theoretical or empirical) of 0 implies no linear dependence of the variables. If you assume a bivariate normal distribution it also implies independence of X and Y.
- A correlation of ± 1 implies a perfect linear dependence between the variables.

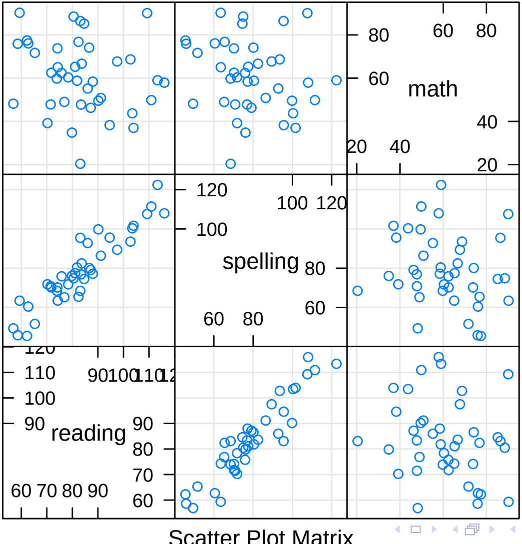
Heuristic Data Generation

```
> library(MASS)
> set.seed(12346)
> cov.mat <- matrix(c(225, 200, 30, 200, 225, 15,
+     30, 15, 225), 3, 3, dimnames = list(c("reading",
+     "spelling", "math"), c("reading", "spelling",
+     "math")))
> studknow <- data.frame(mvrnorm(40, c(80, 78, 64),
+     cov.mat))
> head(studknow)
```

```
    reading  spelling     math
1 103.46649 100.33448 43.75703
2  75.77614  75.87772 62.43045
3  94.60047  95.59099 38.27453
4  56.87628  49.39053 48.18428
5  62.71829  60.47098 76.07416
6 115.98872 107.98283 57.91762
```

Scatterplot matrix of heuristic measures

```
> print(splom(~studknow, aspect = 1, type = c("g",
+     "p")))
```



Computing Pearson r

- We can now compute the statistic for Pearson r across all of our measures with:

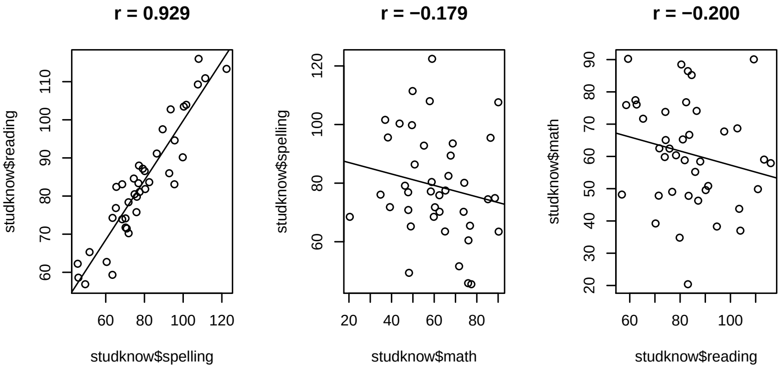
```
> cor(studknow)
```

```
            reading   spelling       math
reading   1.0000000  0.9288218 -0.1995084
spelling  0.9288218  1.0000000 -0.1789046
math     -0.1995084 -0.1789046  1.0000000
```

- The Pearson r also answers the question for us of "How well does a single line represent the bivariate relationship between these two vectors of data?"
- By plotting this, we can see how this is true.

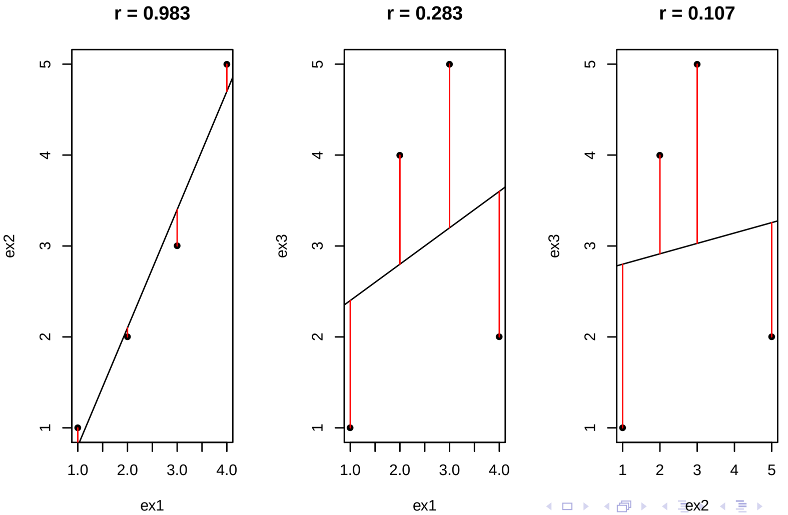
Plotting of bivariate relationships

```
> par(mfrow = c(1, 3))
> plot(studknow$spelling, studknow$reading, main = "r = 0.929")
> abline(lm(studknow$reading ~ studknow$spelling))
> plot(studknow$math, studknow$spelling, main = "r = -0.179")
> abline(lm(studknow$spelling ~ studknow$math))
> plot(studknow$reading, studknow$math, main = "r = -0.200")
> abline(lm(studknow$math ~ studknow$reading))
```



A Single Line Representing Relationships

```
> ex1 <- c(1, 2, 3, 4)
> ex2 <- c(1, 2, 3, 5)
> ex3 <- c(1, 4, 5, 2)
```



The Simple Linear Model

- In linear regression, we use an observed data to formulate a model about a response variable, say `y`, such that `y` is a function of one or more predictors (or covariates) and a residual (or noise) term.
- For data (x_i, y_i), $i = 1, \ldots, n$ the model is written

$$y_i = a + b * x_i + \epsilon_i \quad i = 1, \ldots, n.$$

  That is, $a + b * x$ is the "prediction" part and ϵ is the "noise" part.
- It follows that it is rare that we would actually have perfect prediction in a linear model, hence we need to include our residual term ϵ.

Assumptions for the Residual Term ϵ

- We assume the values for ϵ_i are independent and identically distributed (i.i.d.) normal random variables with mean 0 and (common) variance σ^2, or

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- This assumption is similar to the assumption of a "random sample from a normal distribution" for the one-sample t-test.
- Just because we assume independence and constant variance properties does not make them true. We need to assess these assumptions after fitting any models. (This will be discussed later)

The Least Squares Regression Line

- With correlation, we looked asked the question "How well does a single line represent the relationship between two variables?"
- With regression, we determine *where* to draw that line.
- The line that we fit is the ordinary least squares (OLS) line. We find values for a and b that minimize the squared distances between each actual y x combination and that fitted line.
- Put another way, we want to *minimize* the sum of squares residual by finding values for a and b that produce the smallest

$$SS_{res} = \sum_{i=1}^{n} [y_i - (a + b * x_i)]^2$$

- Notice how the above equation represents the squared distance of each person from their "predicted" score based on a and b.

The `lm` Function in R

- In the `lm` function, we specify the dependent variable as modeled by ($\sim$) the independent variable.

```
> new.data <- data.frame(dv = 1:10, iv = c(1, 3,
+     2, 5, 4, 6, 6, 8, 9, 11))
> summary(m1 <- lm(dv ~ iv, new.data))

Call:
lm(formula = dv ~ iv, data = new.data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1934 -0.5180  0.1160  0.6146  1.0387

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.42541    0.54433   0.782    0.457
iv           0.92265    0.08683  10.626 5.39e-06

Residual standard error: 0.826 on 8 degrees of freedom
Multiple R-squared: 0.9338, Adjusted R-squared: 0.9256
F-statistic: 112.9 on 1 and 8 DF,  p-value: 5.385e-06
```

Interpretation of the Coefficients

- The term labeled `(Intercept)` is really the expected value for the dependent variable when the independent variable is 0 (or all independent variables are 0).
- The term labeled `iv` is the slope for the independent variable, which we conveniently labeled `iv`. This is the expected change in the dependent variable for every one unit change in the independent variable.
- In this case, we expect the dependent variable to go up `0.92265` points for every $+1$ change in `iv`.
- We also think of the slope as "rise over run", or $\frac{\text{rise}}{\text{run}}$.

Statistical Significance in Regression

- The most important hypothesis that we test in regression is whether or not the slope for our predictor variable is statistically significantly different from 0, or $H_0 : b = 0$.
- If we decide not to reject H_0, then we can reduce our original model to $y_i = a + \epsilon_i$. This is the same thing as saying "our predictor variable provides no more information in to describing the variability of the dependent variable than if we just guessed the mean of the dependent variable each time."
- We generally are not interested in testing $H_0 : a = 0$ since it is only examining whether or not the "y-intercept" is statistically significantly different from 0.
- The `t value` for each effect is found by taking the estimate for that effect divided by its standard error. This allows us to test for the probability of that effect (`Pr(>|t|)`) given our df.

Statistical Significance or of the Whole Model

- As opposed to testing the (`Pr(>|t|)`) for an individual effect, we can also test the entire effect of all covariates included in our model.
- In the case of a single predictor model, we test the difference between a trivial model, $y_i = a + \epsilon_i$, and our full model, $y_i = a + b * x + \epsilon_i$.
- For our single predictor model, the F-statistic is the same as the square of the t-statistic.

Practical Significance - R^2

- In the summary of our data, we see that the `Multiple R-squared` is 0.9338. In the case of multiple regression with one predictor, this is the same as the squared correlation between the dependent variable and the covariate.

```
> cor(new.data)^2

          dv        iv
dv 1.0000000 0.9338356
iv 0.9338356 1.0000000
```

- The multiple R^2 is the same thing as the squared correlation between the fitted values and the dependent variable.

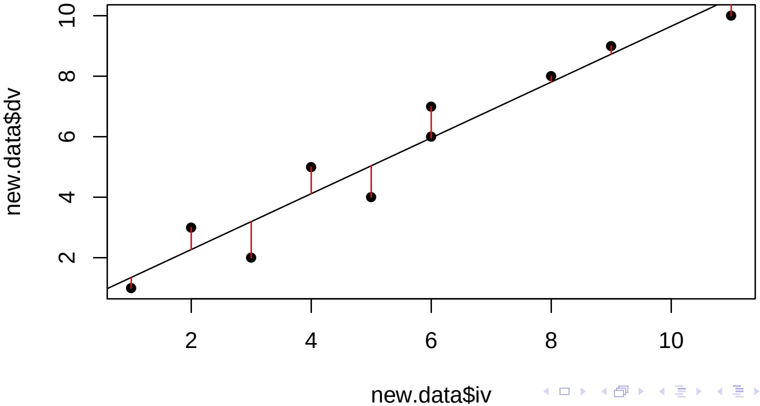
```
> cor(fitted(m1), new.data$dv)^2

[1] 0.9338356
```

- These fitted values are sometimes referred to as $\hat{y}$ and are obtained by computing them from $\hat{y} = \hat{a} + \hat{b} * x$.

Plotting Heuristic Data

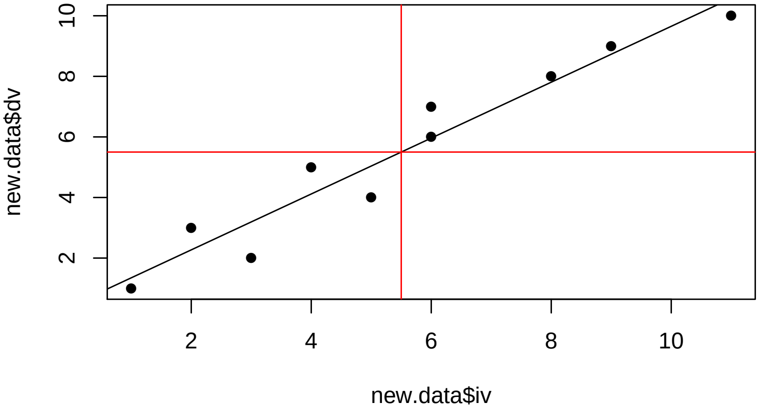
```
> plot(new.data$iv, new.data$dv, pch = 16)
> abline(lm(new.data$dv ~ new.data$iv))
> segments(new.data$iv, fitted(m1), new.data$iv,
+      new.data$dv, col = "red")
```



The Centroid

```
> mean(new.data)

 dv  iv
5.5 5.5
```

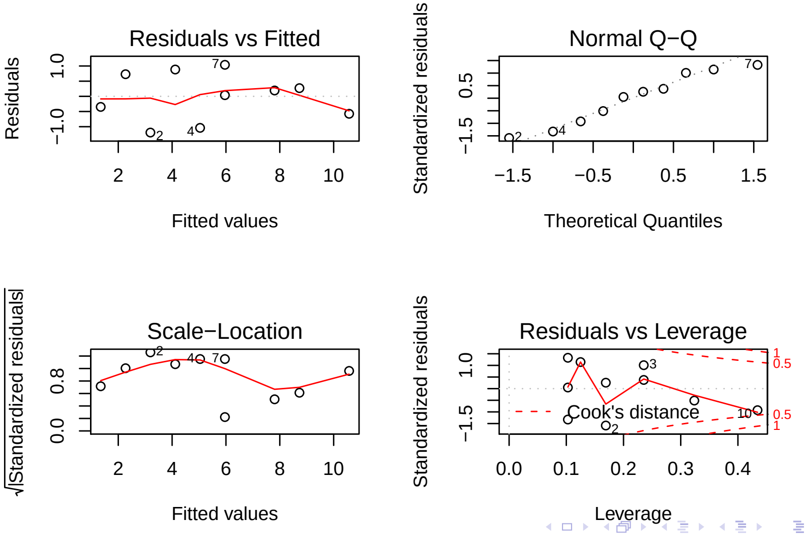


Examining Fitted Values and Residuals

```
> cbind(new.data, fit = m1$fit, resid = m1$resid)
   dv iv       fit       resid
1   1  1  1.348066 -0.34806630
2   2  3  3.193370 -1.19337017
3   3  2  2.270718  0.72928177
4   4  5  5.038674 -1.03867403
5   5  4  4.116022  0.88397790
6   6  6  5.961326  0.03867403
7   7  6  5.961326  1.03867403
8   8  8  7.806630  0.19337017
9   9  9  8.729282  0.27071823
10 10 11 10.574586 -0.57458564
```

Checking Assumptions of the Linear Model

```
> par(mfrow = c(2, 2))
> plot(m1)
```



Identifying Outliers

- We can look at the influence of each individual variable pairs by looking at their influence on the coefficients (a and b) when they are removed.

```
> influence(m1)$coefficients

    (Intercept)            iv
1  -0.192232694  0.0255931102
2  -0.361819678  0.0396732103
3   0.298246735 -0.0368856386
4  -0.150940315  0.0063957761
5   0.193091132 -0.0167420057
6   0.003000572  0.0002381406
7   0.080586778  0.0063957761
8  -0.012085635  0.0064285294
9  -0.039903554  0.0136923961
10  0.237914365 -0.0617230663
```

Homework Part 1

1. Create a dataset with at least 20 people where the correlation between the dv and the iv is at least 0.80. Also, create the data in such a way that the slope coefficient for the iv is 1.0 $\pm$ 0.10.
2. Run a linear regression using `lm` and produce both the output and a scatterplot with the OLS line of best fit. Name this model `m1`.

Homework Part 2

1. Changing only the iv from the data in `m1`, run a new linear modle in which the slope coefficient for the iv is 3.0 ± 0.30.
2. Run a linear regression using `lm` and produce both the output and a scatterplot with the OLS line of best fit. Name this model `m2`.

Homework Part 3

1. Go back to your data from `m1`. In this case, chane two of the values on the iv in such a way that their Cook's Distance is now outside of the confidence intervals (make outliers).
2. Run a linear regression using `lm` and produce both the output and a scatterplot with the OLS line of best fit. Include the printout of the model diagnostics. Name this model `m3`.