Problem A

There have been several recent projects directed at replication of published studies. One called *Psych File Drawer* allows psychologists to upload reports of attempts to replicate studies to a website.

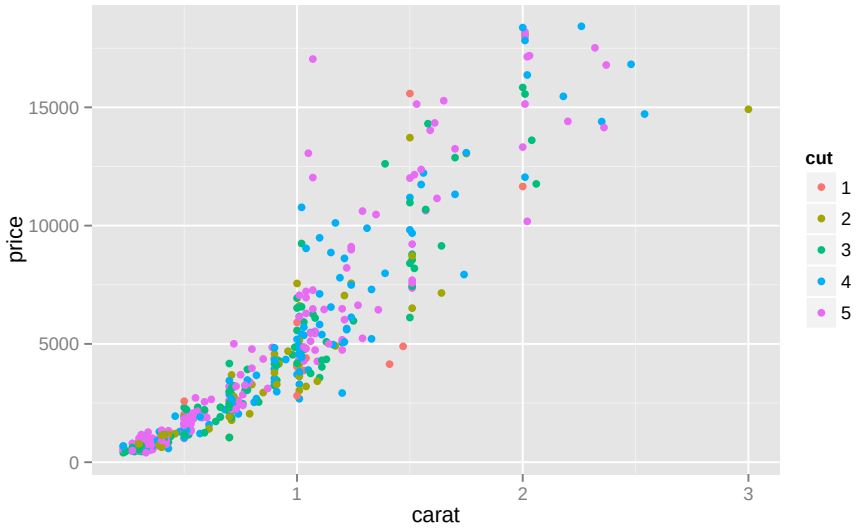
C. Glenn Begley, while a research director at Amgen, attempted to replicate 53 cancer studies. Only 6 of 53 were successfully replicated. A successful replication means that the null hypothesis rejected in the original study was also rejected in the replication at the 5% significance level.

(a) Is there any possible explanation for the common failure to replicate that does not involve dishonest behavior by the authors of the original studies? Briefly explain.

(b) Assume that journals only publish papers with a p-value less than .05. If the null hypothesis were true in 80% of all studies conducted, and the studies all had power equal to 40%, what would be the probability that a paper featured an incorrect rejection of the null hypothesis, given that it was published? In other words, assuming that P(H_0 True) = 0.8, and P(Reject H_0|H_0 True) = 0.05 what is P(H_0 True|RejectH_0)? Recall that *power* is the probability of rejecting H_0 if it is False.

Problem B

The following scatterplot shows 800 diamonds and the relationship between their size (measured in carats), the quality of their cut (measured from a low of 1 to a high of 5) and their price.



While it is clear from the scatterplots that there were some non-linear trends in the data, it turns out that with some transformations, a linear model becomes more reasonable. Consider two models for the price of diamonds, Model 1 (`m1`) and Model 2 (`m2`), detailed on the back.

(a) Please construct a confidence interval for the true slope associated with cut quality under Model 1. Please use the actual numbers in the construction, but there is no need to go through the arithmetic to get the final interval.

(b) How would you interpret the above confidence interval?

(c) Why do you think that the slope for cut quality is negative in Model 1 and positive in Model 2?

(d) Once you have a model that you're confident does a good job of predicting diamond prices, you plan to use it in a small business operation of buying and reselling diamonds. Say that diamonds come on the market with their cut, carat, and asking price all provided. Which diamonds would you buy? At what price would you resell them?

```
m4 <- lm(logprice ~ cut, data = diamonds)
summary(m4)$coef
```

```
##                Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)  8.10992538 0.16119672 50.310734 2.251424e-197
## cut         -0.06720391 0.04009422 -1.676149  9.433655e-02
```

```
m5 <- lm(logprice ~ cut + sqrtcarat, data = diamonds)
summary(m5)$coef
```

```
##               Estimate Std. Error  t value      Pr(>|t|)
## (Intercept) 4.44995518 0.07038096 63.22669 8.499085e-240
## cut         0.03123363 0.01205615  2.59068  9.860013e-03
## sqrtcarat   3.75322089 0.05263850 71.30182 3.641780e-263
```

Problem C

Using your general knowledge about the world, think about the relationship between these variables:

`speed` of a bicyclist

`steepness` of the road (a numerical continuous variable measured by the grade – aka rise over run). 0 means flat, positive values mean uphill, negative values mean downhill.

`fitness` of the rider, a categorical variable with two levels: `average` and `athletic`

Sketch out a graph of speed versus steepness (be sure to label your axes and lines) for reasonable models (note: we don't yet have data) of each of these forms:

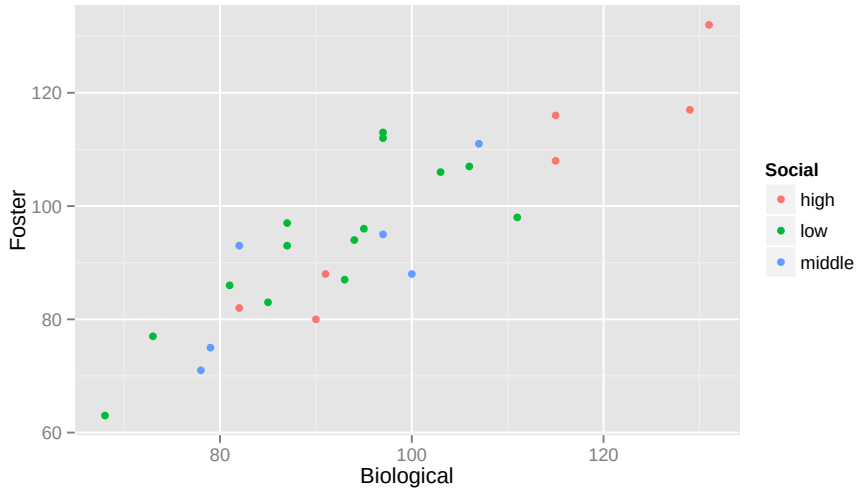
1. Model 1: $\widehat{speed} \sim b_0 + b_1 \times steepness$

2. Model 2: $\widehat{speed} \sim b_0 + b_1 \times steepness + b_2 \times fitness$

3. If you wanted to estimate the parameters for Model 2, you might consider collecting some data. What would you select as your observational unit, i.e. what would constitute a single case, i.e what is represented in each row of the data set? Be as specific as you can.

Problem D

Recall from one of the labs the twins data set collected by Cyril Burt. As a reminder, in this study he tracked down identical twins that were separated at birth: one child was raised in the home of their biological parents and the other in a foster home. In an attempt to answer the question of whether intelligence is result of nature or nurture, both children were given IQ tests.



In that lab we fit the simple model to predict the foster twin's IQ based on the IQ of the biological twin.

```
m1 <- lm(Foster ~ Biological, data = twins)
summary(m1)$coef

##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 9.207599 9.29989643 0.9900754 3.316237e-01
## Biological  0.901436 0.09633286 9.3575128 1.203600e-09
```

1. This data set is used as evidence in the argument of whether intelligence is a result of nature or nurture. In their most absolute rendition, what would each side in this argument posit as the true slope that relates the biological IQ to the foster IQ? Which end of the argument does this model more strongly support?

2. We fit a model (see below) that incorporates additional information about the socio-economic status of the biological family as well as an interaction term between that status and the IQ of the biological twin. The result is a model with multiple intercepts and multiple slopes. What is the slope of the line that corresponds to high social class?

3. Do these results change our sense of how IQ might be related to genes versus the environment?

```
m2 <- lm(Foster ~ Biological + Social + Biological:Social, data = twins)
summary(m2)$coef
```

```
##                              Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)             -1.872043663 17.8082644 -0.1051222 9.172765e-01
## Biological               0.977562159  0.1631923  5.9902480 6.042489e-06
## Sociallow                9.076653525 24.4487043  0.3712529 7.141683e-01
## Socialmiddle             2.688068048 31.6041780  0.0850542 9.330240e-01
## Biological:Sociallow    -0.029139714  0.2445802 -0.1191418 9.062954e-01
## Biological:Socialmiddle -0.004995209  0.3295253 -0.0151588 9.880486e-01
```