# Explanation in regression

ST552 Lecture 12

Charlotte Wickham

Feb 01, 2016

# Sampling

Our model says, we have fixed our $X$ and our responses are generated according to the model

$$Y = X\beta + \epsilon \quad \epsilon \sim N(0, \sigma^2 I)$$

but how does this relate to real life data?
**Designed experiments**: We fix $X$ and let nature generate the repsonses according to the model. We only observe a finite number of observations and our inference tells us about the $\beta$ underlying the natural generating process.
**Observational studies**: A population of responses exists for each unique $X$. We observe a sample from the populations and use our estimate to make inferences on the population value of $\beta$. Generally, we like simple random samples and sample much smaller than the population (or use finite population corrections).

# Sampling cont.

**Complete population**: Permutation tests give some meaning to the p-value for the sample at hand (more later). Or use regression just as a descriptive tool for the sample at hand. Or imagine parallel alternative worlds.

**Random** $X$: some set up regression conditional on $X$. Can show many of the same properties, i.e. $E(\hat{\beta}|X) = \beta)$, but must add crucial assumption that $X$ and $\epsilon$ are independent.

```r
library(faraway)
data(gala, package = "faraway")
lmod <- lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
  data = gala)
sumary(lmod)
```

```
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  7.068221  19.154198  0.3690 0.7153508
## Area        -0.023938   0.022422 -1.0676 0.2963180
## Elevation    0.319465   0.053663  5.9532 3.823e-06
## Nearest      0.009144   1.054136  0.0087 0.9931506
## Scruz       -0.240524   0.215402 -1.1166 0.2752082
## Adjacent    -0.074805   0.017700 -4.2262 0.0002971
##
## n = 30, p = 6, Residual SE = 60.97519, R-Squared = 0.77
```

What is the meaning of $\hat{\beta}_{\text{Elevation}} = 0.32$?

# Naive (**wrong**) interpretation

> *A unit increase in* $x_1$ *will produce a change of* $\beta_1$ *in the response*

Where does this come from? Compare $(y_i|X_{i1} = x_1)$ to $(y_i|X_{i1} = x_1 + 1)$
Or compare $E(y_i|X_{i1} = x_1)$ to $E(y_i|X_{i1} = x_1 + 1)$

> *A unit increase in* $x_1$ *is associated with a change of* $\beta_1$ *in the mean response*

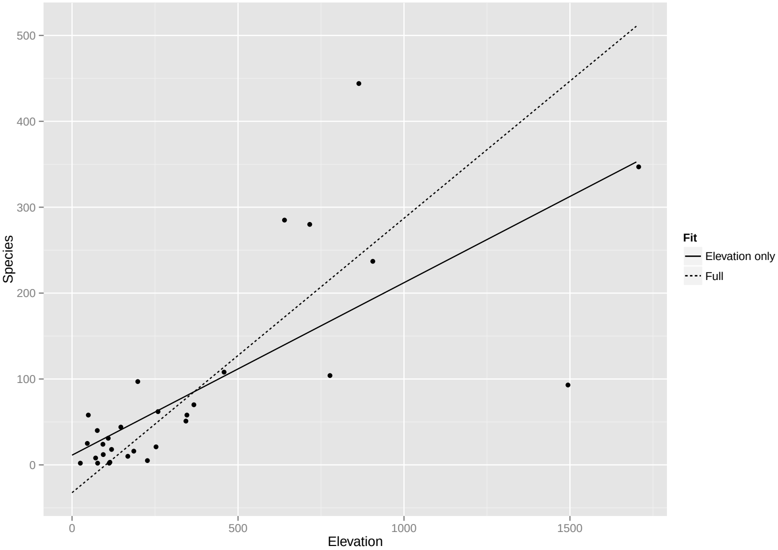
> *Compare two islands where the elevation of one island is one meter higher than the first. On average the second island has 0.32 species more than the first.*

# More accurate interpretation

You must be specific about what else is in the model, because the meaning of $\beta_1$ is different in the following models

$$y_i = \beta_0 + \beta_1 \text{Elevation}_i + \epsilon_i$$

$$y_i = \beta_0 + \beta_1 \text{Elevation}_i + \beta_2 \text{Area}_i + \beta_3 \text{Nearest}_i + \beta_4 \text{Scruz} + \beta_5 \text{Adjacent} + \epsilon_i$$


Species
Elevation
0
100
200
300
400
500
0
500
1000
1500
Fit
Elevation only
Full

# Better

*A unit increase in* $x_1$ *with the other (named) predictors held constant is associated with a change of* $\beta_1$ *in the response*

*Compare two islands with the same area, distance to nearest island, distance from Santa Cruz island, and area of the adjacent island, but where the elevation of one island is one meter higher than the first. On average the second island has 0.32 species more than the first.*

But these are purely hypothetical islands. We can't change any of these properties, let alone one without the others.

# Causal inference

```r
library(faraway)
data(newhamp, package = "faraway")
newhamp$handcount <- ifelse(newhamp$votesys == "H", 1, 0)
lmod <- lm(pObama ~ handcount,
  data = newhamp)
sumary(lmod)
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 0.3525171  0.0051728 68.1480 < 2.2e-16
## handcount   0.0424871  0.0085091  4.9932 1.059e-06
##
## n = 276, p = 2, Residual SE = 0.06823, R-Squared = 0.08
```

$$\text{Proportion for Obama}_i = y_i = \beta_0 + \beta_1 \text{handcount}_i + \epsilon_i$$

What is the meaning of $\hat{\beta}_1 = 0.04$?

# Your Turn: Indicator Variables

$$\text{handcount}_i = \begin{cases} 1, & \text{ward i counted votes by hand} \\ 0, & \text{otherwise (ward i counted by machine)} \end{cases}$$

Find $E(y_i|\text{handcount}_i = 1)$, and $E(y_i|\text{handcount}_i = 0)$.

What is $E(y_i|\text{handcount}_i = 1) - E(y_i|\text{handcount}_i = 0)$?

Explain in words in context of the problem.

# Equivalent to an equal variance two sample t-test

```
t.test(pObama ~ handcount, data = newhamp, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  pObama by handcount
## t = -4.9932, df = 274, p-value = 1.059e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05923854 -0.02573565
## sample estimates:
## mean in group 0 mean in group 1
##       0.3525171       0.3950042
```

*The wards with handcounting had on average a proportion of votes for Obama 4% higher than the wards with digital counts.*

Can me make the causal conclusion?

*Handcounting increased the proportion of votes for Obama by 4%*

# Causality

A causal effect is the difference between outcomes where action was taken or not.
Let $T = 1$ for the treatment and $T = 0$ be the control. Then $y_i^1$ be the observed response for subject $i$ under the control and $y_i^0$ be the observed response for subject $i$ under the treatment.
We want to know:

$$\delta_i = y_i^1 - y_i^0$$

**The fundemental problem:** we can only observe one of $(y_i^1, y_i^0)$.

# Ways to proceed

**Randomized experiment** We control which observation we make. We randomize units to treatments. On average, confounders will be balanced across treatment groups. But, more importantly randomization provides a complete basis for inference about the average treatment effect. Causal inference in justified without further assumptions.
**Close substitution** We argue that we can observe things that are very close to $y_i^1$ and $y_i^0$.
**Matching** We match observations based in similar covariates to provide a fair comparison.
**Statistical adjustment** We know about confounders and model their effects to remove their contribution.
See Faraway Chapter 5 for covariate adjustment and matching for the voting data.