Lesson: Motivating statistical inference in the linear regression context

Module: Inference in Linear Regression

Photo by Peter Pryharski on Unsplash

Marketing data

- `youtube`: the advertising budget allocated to YouTube. Measured in thousands of dollars;
- `facebook`: the advertising budget allocated to Facebook. Measured in thousands of dollars; and
- `newspaper`: the advertising budget allocated to a local newspaper. Measured in thousands of dollars.
- `sales`: the value in the ith row of the sales column is a measurement of the sales (in thousands of units) for product P for company i.

Photo by Denis Cherkashin on Unsplash

```
Call:
lm(formula = sales ~ facebook + youtube + newspaper, data = marketing)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.5932  -1.0690   0.2902   1.4272   3.3951 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.526667   0.374290   9.422   <2e-16 ***
facebook     0.188530   0.008611  21.893   <2e-16 ***
youtube      0.045765   0.001395  32.809   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.023 on 196 degrees of freedom
Multiple R-squared:  0.8972,	Adjusted R-squared:  0.8956 
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Photo by Denis Cherkashin on Unsplash

```
Call:
lm(formula = sales ~ facebook + youtube + newspaper, data = marketing)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.5932  -1.0690   0.2902   1.4272   3.3951 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.526667   0.374290   9.422   <2e-16 ***
facebook     0.188530   0.008611  21.893   <2e-16 ***
youtube      0.045765   0.001395  32.809   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.023 on 196 degrees of freedom
Multiple R-squared:  0.8972,	Adjusted R-squared:  0.8956 
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Photo by Denis Cherkashin on Unsplash

Is there evidence to suggest that the parameter associated with `newspaper` is *statistically different from zero*, i.e., is the `newspaper` parameter **statistically significant**?

Independent of statistical significance, is there reason to believe that the parameter associated with `newspaper` is *meaningfully different from zero,* i.e., is the `newspaper` parameter **practically significant**?

Photo by Denis Cherkashin on Unsplash

	youtube	facebook	newspaper	sales
Company #1	276,12	45,36	83,04	Y_1
Company #2	53,40	47,16	54,12	Y_2
...	...	...	...	...
Company #n	278,52	10,32	10,44	Y_n

$$\hat{\beta}_I \quad \hat{\beta}_{II} \quad \hat{\beta}_{III} \quad \dots \hat{\beta}_{IV}$$

Photo by Denis Cherkashin on Unsplash

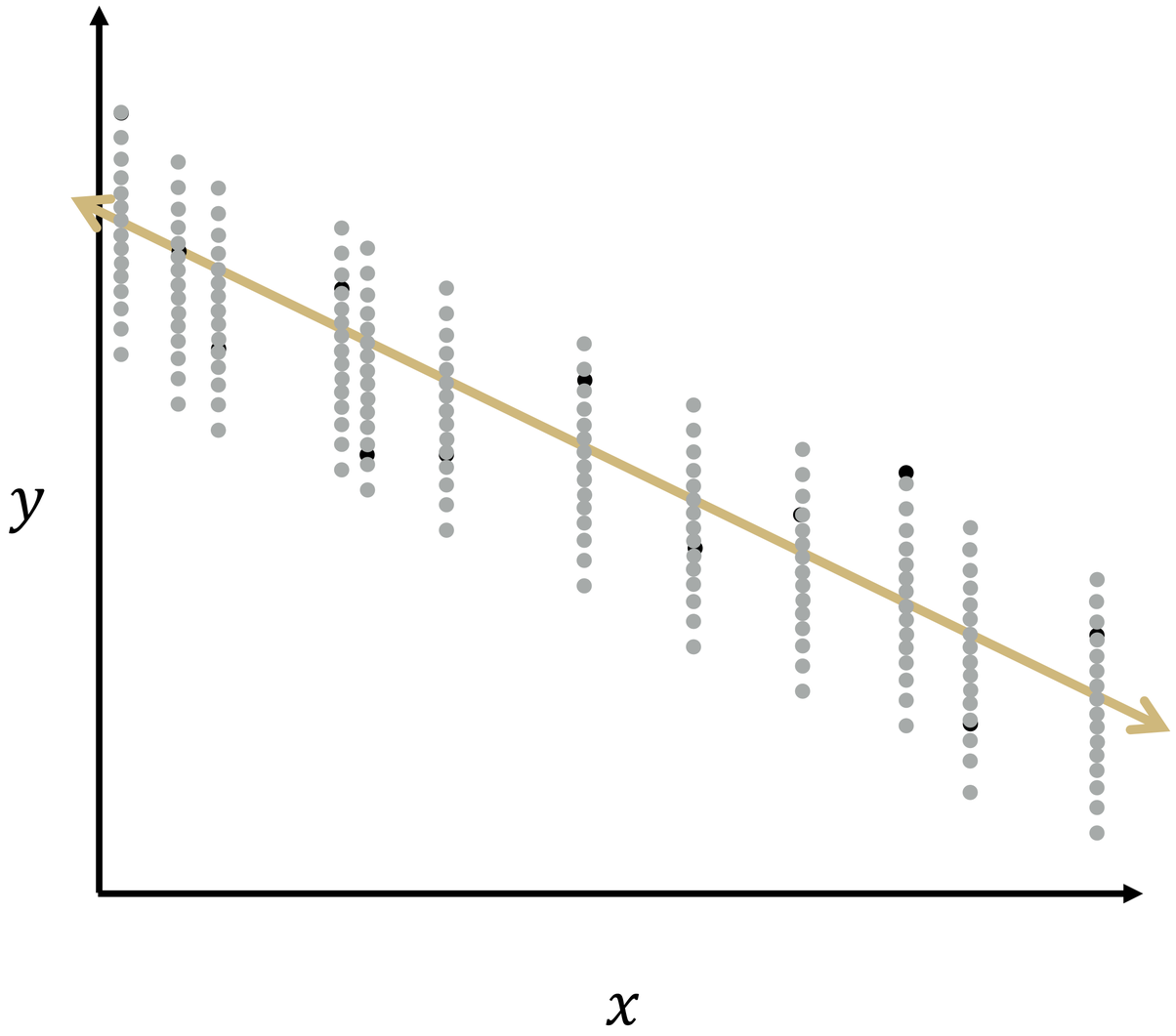
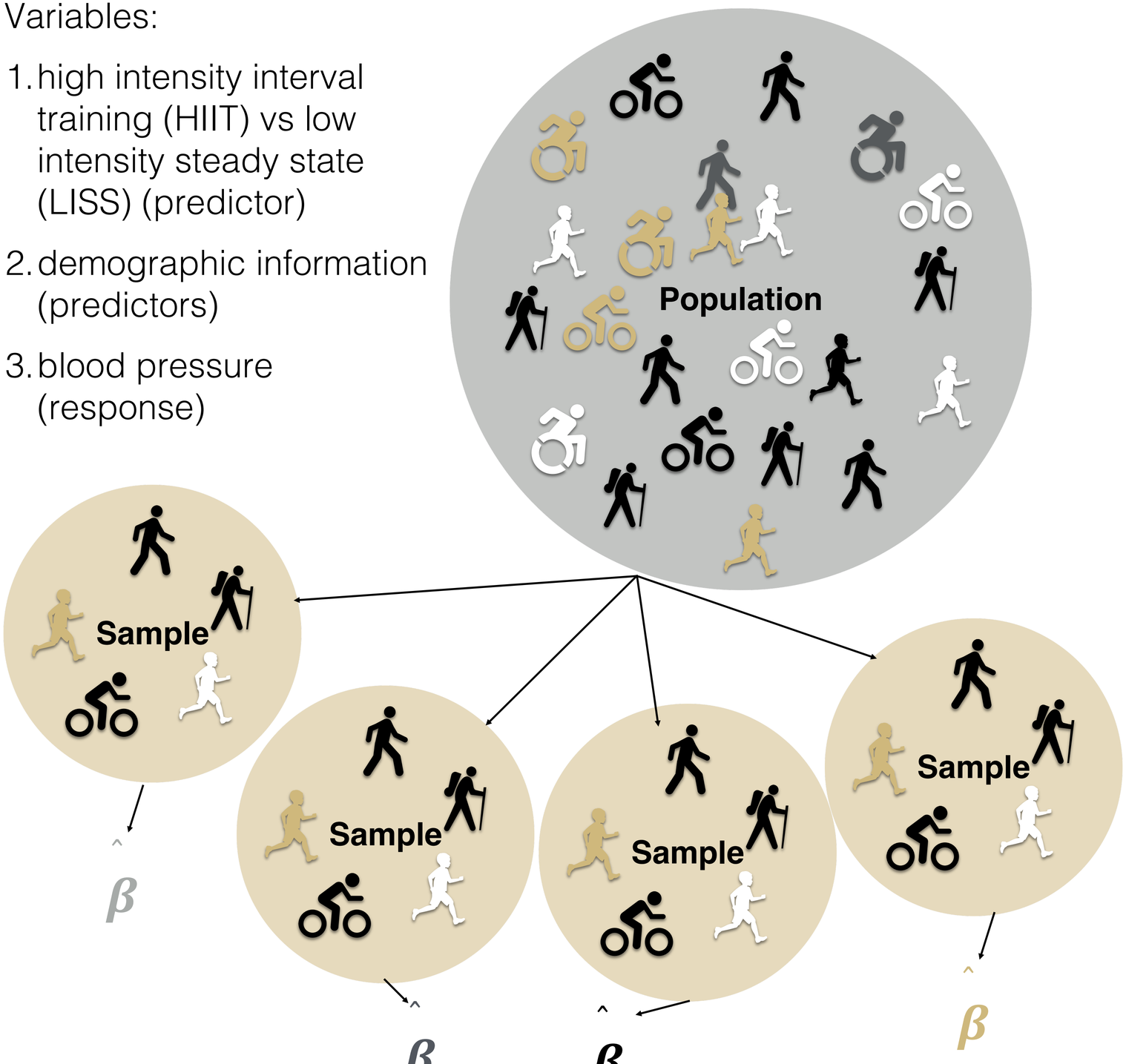


Photo by Denis Cherkashin on Unsplash

Variables:

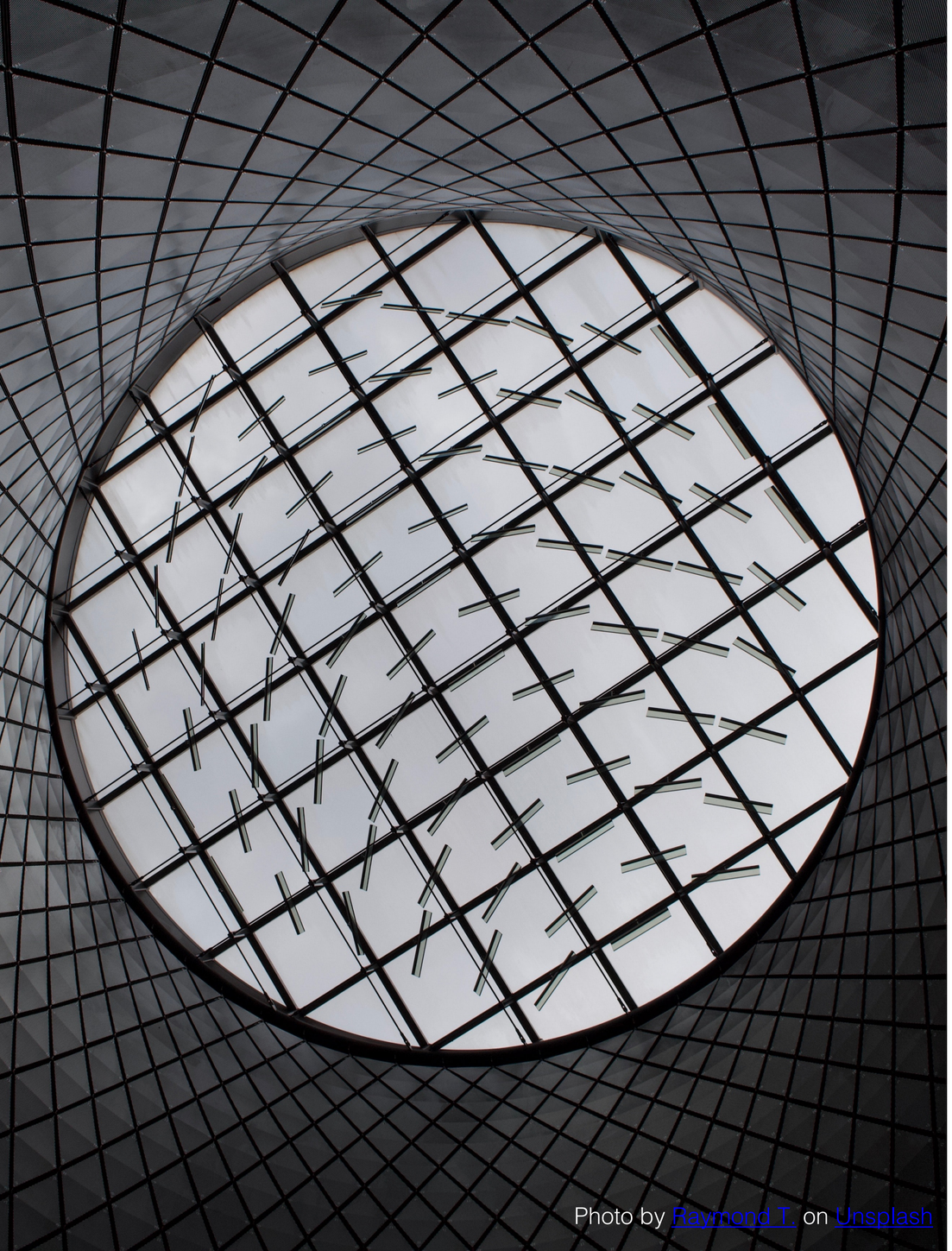
1. high intensity interval training (HIIT) vs low intensity steady state (LISS) (predictor)
2. demographic information (predictors)
3. blood pressure (response)



Lesson: The sampling distribution of the least squares estimator

Module: Inference in Linear Regression

Photo by Peter Pryharski on Unsplash



Definition: The *sampling distribution* of the least squares estimator, $\hat{\boldsymbol{\beta}}$, is defined as the probability distribution of $\hat{\boldsymbol{\beta}}$, treated as a random variable across random samples of size n.

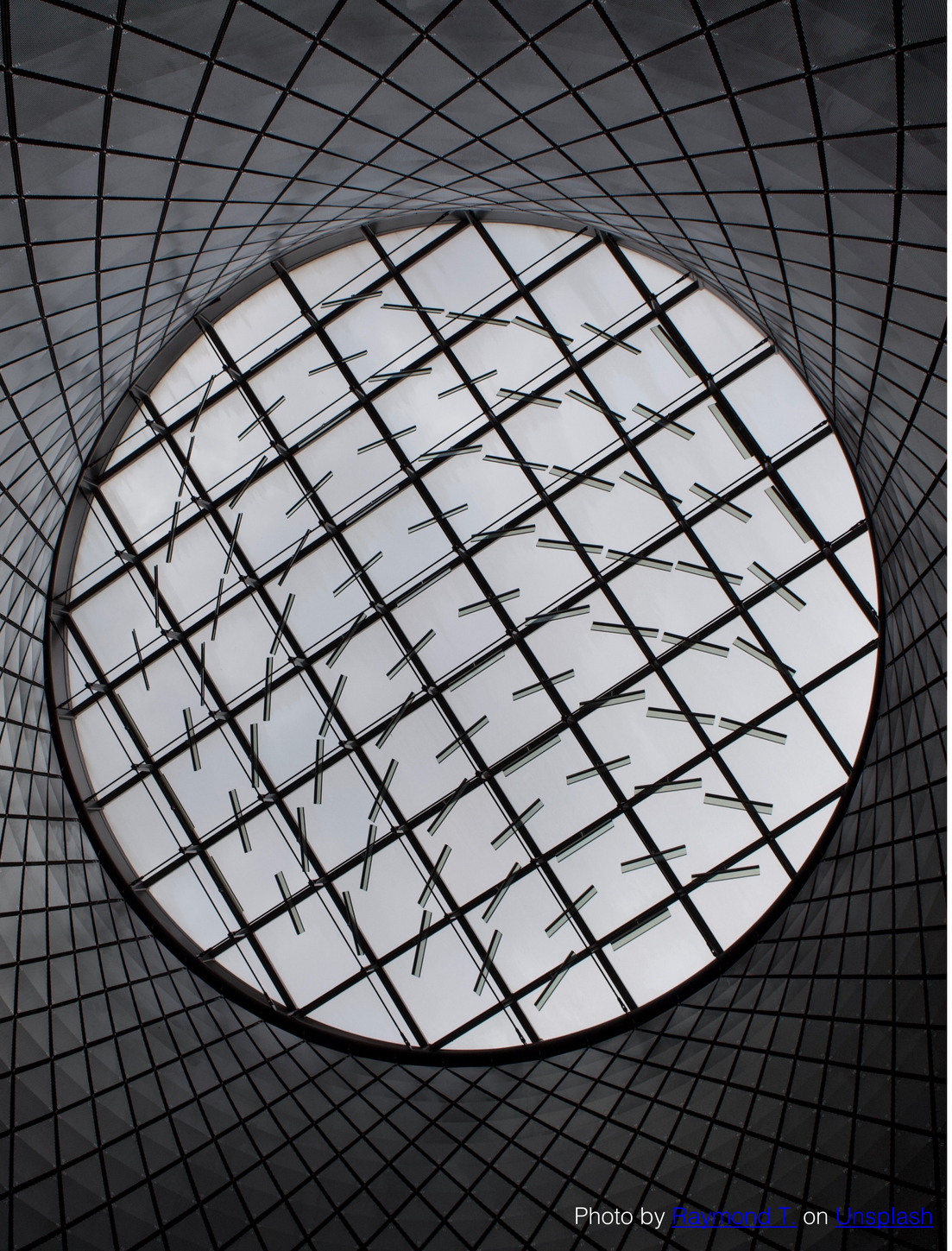
With the sampling distribution, we could compute:

1. ***Statistical Significance.*** Test particular claims about regression parameters, for example:

$$H_0: \beta_j = 0 vs \beta_j \neq 0,$$

for some $j = 1, \ldots, p$.

2. ***Confidence intervals.*** Find ***interval estimates*** of regression parameters.



Note that $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y}$ is a linear

constant

combination of normal random variables. Thus:

Or, individually:

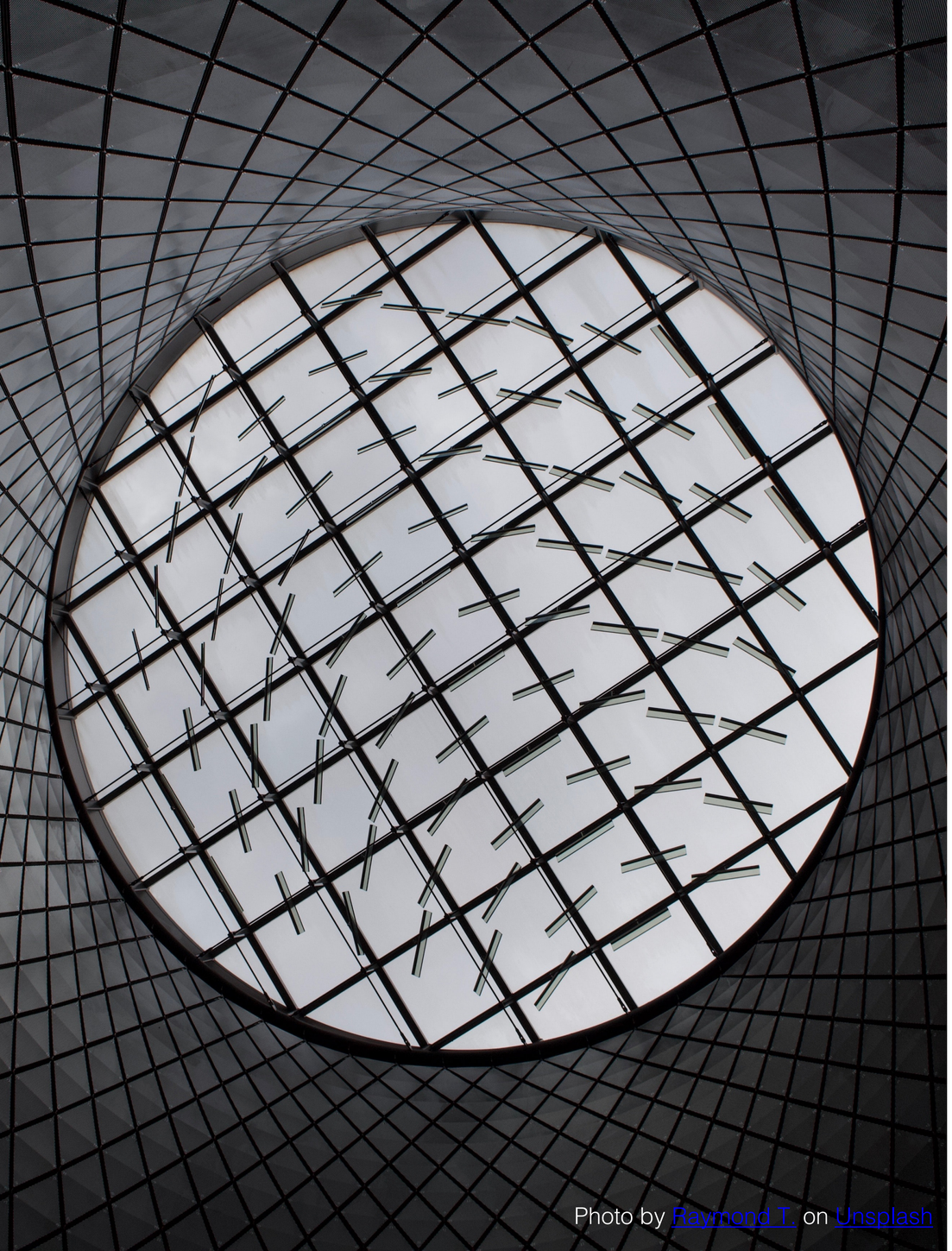


Photo by Raymond T. on Unsplash

1. Show that: $E\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}$.

2. Show that: $Var\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2(X^T X)^{-1}$.

Lesson: t-tests for individual regression parameters

Module: Inference in Linear Regression

Photo by Peter Pryharski on Unsplash

Suppose that we want to test:

$$H_0: \beta_j = c vs H_1: \beta_j \neq c, c \in R$$

After specifying a significance level α, we need to calculate a test statistic.

General form of test statistics:

Test statistic for above hypotheses:

where:

```
Call:
lm(formula = sales ~ facebook + youtube + newspaper, data = marketing)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.5932  -1.0690   0.2902   1.4272   3.3951 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.526667   0.374290   9.422   <2e-16 ***
facebook     0.188530   0.008611  21.893   <2e-16 ***
youtube      0.045765   0.001395  32.809   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.023 on 196 degrees of freedom
Multiple R-squared:  0.8972,	Adjusted R-squared:  0.8956 
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Photo by Denis Cherkashin on Unsplash

Lesson: Motivating the F-test: multiple statistical comparisons

Module: Inference in Linear Regression

Photo by Peter Pryharski on Unsplash

Photo by Tetiana Bykovets on Unsplash

Does chocolate everyday cause weight loss?

Photo by Tetiana Bykovets on Unsplash

Suppose that we conduct 18 hypothesis tests at significance level $\alpha = 0.05$, i.e., the probability of type I error is 0.05 for *any individual test.*

What is the probability of a type I error on *at least one* of these 10 tests?

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.1764134  0.1459311  -1.209  0.23251
V1          -0.0522310  0.1227266  -0.426  0.67227
V2           0.1631277  0.1411252   1.156  0.25332
V3           0.4423084  0.1364300   3.242  0.00214 **
V4          -0.0146449  0.1428449  -0.103  0.91876
V5           0.0684142  0.1480405   0.462  0.64603
V6          -0.0162695  0.1250285  -0.130  0.89700
V7           0.1392489  0.1326155   1.050  0.29886
V8           0.2811918  0.1379869   2.038  0.04699 *
V9          -0.1047692  0.1417033  -0.739  0.46322
V10         -0.0594510  0.1535042  -0.387  0.70022
V11          0.2120031  0.1456582   1.455  0.15191
V12         -0.2711908  0.1427382  -1.900  0.06334 .
V13         -0.0101344  0.1563509  -0.065  0.94858
V14          0.3019595  0.1373681   2.198  0.03269 *
V15          0.2073799  0.1199333   1.729  0.09008 .
V16         -0.0441717  0.1275577  -0.346  0.73061
V17          0.3687316  0.1574882   2.341  0.02333 *
V18         -0.2160815  0.1451869  -1.488  0.14308
V19         -0.1489516  0.1246523  -1.195  0.23786
V20         -0.0297282  0.1026911  -0.289  0.77343
V21         -0.0953399  0.1217123  -0.783  0.43721
V22          0.1259770  0.1402307   0.898  0.37339
V23         -0.0857774  0.1541746  -0.556  0.58049
V24          0.0088037  0.1333081   0.066  0.94761
V25         -0.0514570  0.1396584  -0.368  0.71413
V26          0.0460939  0.1516710   0.304  0.76249
V27         -0.0190201  0.1500545  -0.127  0.89965
V28          0.0713307  0.1344797   0.530  0.59822
V29          0.2495248  0.1675303   1.489  0.14278
V30         -0.1306864  0.1450446  -0.901  0.37199
V31         -0.1687308  0.1412891  -1.194  0.23814
V32          0.0117419  0.1698927   0.069  0.94518
V33          0.0973685  0.1314703   0.741  0.46246
V34          0.0735950  0.1263718   0.582  0.56299
V35         -0.3847197  0.1322300  -2.909  0.00543 **
V36          0.0006584  0.1299929   0.005  0.99598
V37          0.1975057  0.1192332   1.656  0.10402
V38         -0.1989029  0.1490083  -1.335  0.18810
V39          0.0872970  0.1322515   0.660  0.51229
V40         -0.1080814  0.1435668  -0.753  0.45515
V41         -0.2389443  0.1240537  -1.926  0.05990 .
V42          0.1359314  0.1767643   0.769  0.44559
V43         -0.1571023  0.1405136  -1.118  0.26899
V44          0.1117499  0.1289166   0.867  0.39026
V45          0.2095581  0.1356878   1.544  0.12892
V46         -0.0187331  0.1313402  -0.143  0.88717
V47          0.0316976  0.1451895   0.218  0.82809
V48         -0.2176497  0.1482230  -1.468  0.14839
V49         -0.2666675  0.1285305  -2.075  0.04328 *
V50         -0.0859325  0.1527803  -0.562  0.57637
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p} + \varepsilon_i$$

Photo by Providence Doucet on Unsplash

So, we need a way to control the overall type I error rate when we need to test whether to drop several predictors from a model.

We do that with a ***simultaneous test***: the F-test.

Lesson: The F-test

Module: Inference in Linear Regression

Photo by Peter Pryharski on Unsplash

Photo by Providence Doucet on Unsplash

Prerequisite to the F distribution: ***the Chi-Squared (χ^2) distribution.***

The χ^2 distribution is a special case of the gamma distribution:

$$\chi^2(n) \equiv \Gamma(\frac{n}{2}, \frac{1}{2})$$

The $\chi^2(n)$ distribution arises in statistics in the following way: Let $Z_1, \ldots, Z_n \sim N(0,1)$. Then:

Photo by Providence Doucet on Unsplash

The F-distribution: Let $X_1 \sim \chi^2(d_1)$ and $X_2 \sim \chi^2(d_2)$. A random variable, X, is said to have an *F* distribution, with parameters d_1 and d_2, if:

Photo by Providence Doucet on Unsplash

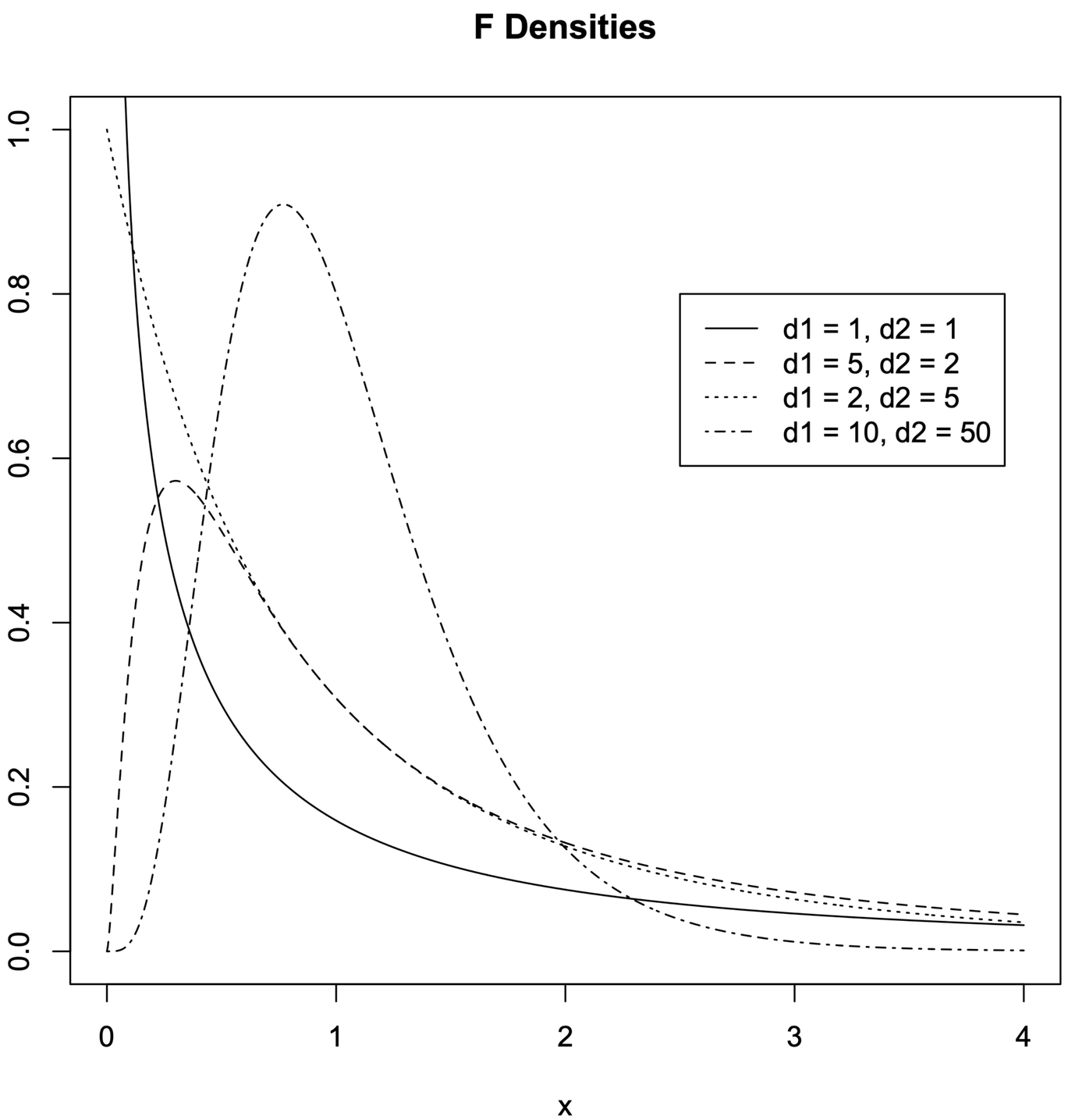

F Densities
d1 = 1, d2 = 1
d1 = 5, d2 = 2
d1 = 2, d2 = 5
d1 = 10, d2 = 50
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
x

Photo by Providence Doucet on Unsplash

Consider two different models, a ***full*** model Ω:

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p} + \varepsilon_i$$

and a ***reduced*** model ω:

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_q x_{i,q} + \varepsilon_i,$$

where $q < p$.

We can consider testing the hypotheses:

H_0:

H_1:

Photo by Providence Doucet on Unsplash

Each model has its own RSS. We use these to construct a test statistic (that ends up having an F-distribution):

$$F = \frac{(RSS_\omega - RSS_\Omega)/(p-q)}{RSS_\Omega/(n-(p+1))}$$

$$\sim F_{p-q,n-(p+1)}$$

Special case (Full F-test):

$$\omega: Y_i = \beta_0 + \varepsilon_i$$

Ω: some larger model

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.1764134  0.1459311  -1.209  0.23251
V1          -0.0522310  0.1227266  -0.426  0.67227
V2           0.1631277  0.1411252   1.156  0.25332
V3           0.4423084  0.1364300   3.242  0.00214 **
V4          -0.0146449  0.1428449  -0.103  0.91876
V5           0.0684142  0.1480405   0.462  0.64603
V6          -0.0162695  0.1250285  -0.130  0.89700
V7           0.1392489  0.1326155   1.050  0.29886
V8           0.2811918  0.1379869   2.038  0.04699 *
V9          -0.1047692  0.1417033  -0.739  0.46322
V10         -0.0594510  0.1535042  -0.387  0.70022
V11          0.2120031  0.1456582   1.455  0.15191
V12         -0.2711908  0.1427382  -1.900  0.06334 .
V13         -0.0101344  0.1563509  -0.065  0.94858
V14          0.3019595  0.1373681   2.198  0.03269 *
V15          0.2073799  0.1199333   1.729  0.09008 .
V16         -0.0441717  0.1275577  -0.346  0.73061
V17          0.3687316  0.1574882   2.341  0.02333 *
V18         -0.2160815  0.1451869  -1.488  0.14308
V19         -0.1489516  0.1246523  -1.195  0.23786
V20         -0.0297282  0.1026911  -0.289  0.77343
V21         -0.0953399  0.1217123  -0.783  0.43721
V22          0.1259770  0.1402307   0.898  0.37339
V23         -0.0857774  0.1541746  -0.556  0.58049
V24          0.0088037  0.1333081   0.066  0.94761
V25         -0.0514570  0.1396584  -0.368  0.71413
V26          0.0460939  0.1516710   0.304  0.76249
V27         -0.0190201  0.1500545  -0.127  0.89965
V28          0.0713307  0.1344797   0.530  0.59822
V29          0.2495248  0.1675303   1.489  0.14278
V30         -0.1306864  0.1450446  -0.901  0.37199
V31         -0.1687308  0.1412891  -1.194  0.23814
V32          0.0117419  0.1698927   0.069  0.94518
V33          0.0973685  0.1314703   0.741  0.46246
V34          0.0735950  0.1263718   0.582  0.56299
V35         -0.3847197  0.1322300  -2.909  0.00543 **
V36          0.0006584  0.1299929   0.005  0.99598
V37          0.1975057  0.1192332   1.656  0.10402
V38         -0.1989029  0.1490083  -1.335  0.18810
V39          0.0872970  0.1322515   0.660  0.51229
V40         -0.1080814  0.1435668  -0.753  0.45515
V41         -0.2389443  0.1240537  -1.926  0.05990 .
V42          0.1359314  0.1767643   0.769  0.44559
V43         -0.1571023  0.1405136  -1.118  0.26899
V44          0.1117499  0.1289166   0.867  0.39026
V45          0.2095581  0.1356878   1.544  0.12892
V46         -0.0187331  0.1313402  -0.143  0.88717
V47          0.0316976  0.1451895   0.218  0.82809
V48         -0.2176497  0.1482230  -1.468  0.14839
V49         -0.2666675  0.1285305  -2.075  0.04328 *
V50         -0.0859325  0.1527803  -0.562  0.57637
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9437 on 49 degrees of freedom
Multiple R-squared:  0.5505,    Adjusted R-squared:  0.0919
F-statistic:   1.2 on 50 and 49 DF,  p-value: 0.2619
```

Photo by Providence Doucet on Unsplash

To run a partial F-test in R, i.e., the test

$$H_0: \beta_i = \beta_j = \ldots = \beta_k = 0$$

$$H_1: some \beta_l \neq 0,$$

we fit both models, say lmod_full and lmod_reduced, and then use the anova() function:

```
anova(lmod_red, lmod_full)
```

Lesson: Confidence intervals in the regression context

Module: Inference in Linear Regression

Photo by Peter Pryharski on Unsplash

Recalling the distribution of our least squares estimator from earlier, we can derive the $(1-\alpha)\times 100\%$ confidence interval to be:

In R, `confint(lm_data)` will produce confidence intervals (at the 95% level) for all parameters in the model `lm_data`.

Suppose for now that we have a model $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ that fits the data well. We'd like to know the average response at a new set of predictors $\mathbf{x}^\star = (1, x_1^\star, \ldots, x_p^\star)$. The point estimate for this average response is:

But what about an interval estimate?

For our confidence interval, we'll need:

$$Var(\mathbf{x}^{\star}\hat{\beta}) =$$

Thus, a confidence interval for the mean response for a given predictor $\mathbf{x}^{\star}$ is:

Photo by Jr Korpa on Unsplash

Confidence interval for the mean response in R:

predict(lm_data, new = $\mathbf{x}^\star$, interval = “confidence”)

Example: Suppose that sales of a product in a particular market (y) are well predicted by YouTube (x_1), Facebook (x_2), and newspaper advertising budgets (x_3). Suppose that a particular new market opens, with advertising characteristics $x_1^\star, x_2^\star, and x_3^\star$.

Suppose we'd like to know the ***average sales in a market with advertising budgets*** $x_1^\star, x_2^\star, and x_3^\star$. The point estimate is:

The interval estimate is:

Recall that prediction is one of the main purposes of using regression models. Suppose for now that we have a model $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ that fits the data well (we'll say more later about checking wither the model fits the data).

Given a new set of predictors, $\mathbf{x}^\star = (1, x_1^\star, \ldots, x_p^\star)^T$, the estimated average value of the response is:

Suppose we'd like to know the *average sales in a market with advertising budgets* $x_1^\star, x_2^\star, and x_3^\star$. The point estimate is:

But what about an interval estimate?

Prediction

The predicted response on the previous slide is a *point estimator*. What is the appropriate *interval estimator*? There are two different cases:

1. **Prediction interval** for a *future observation*.

2. **Confidence interval** for the *mean response*.

Prediction

Example: Suppose that sales of a product in a particular market (y) are well predicted by YouTube (x_1), Facebook (x_2), and newspaper advertising budgets (x_3). Suppose that a particular new market opens, with advertising characteristics $x_1^\star, x_2^\star, and x_3^\star$.

1. *Since $E(\boldsymbol{\varepsilon}) = (0)$, a point estimate of that market's sales would be _____. But in constructing a **prediction interval**, we need the variance of this estimator, which would include $Var(\boldsymbol{\varepsilon})$.*

Prediction

Example: Suppose that sales of a product in a particular market (y) are well predicted by YouTube (x_1), Facebook (x_2), and newspaper advertising budgets (x_3).

1. In predicting a future value using an interval, we are estimating a *random variable*.

2. In estimating the *average* sales in a market with particular advertising characteristics, we are estimating an *unknown fixed parameter*.

Prediction

The $(1-\alpha)\times 100\%$ *prediction interval* for a future value of *y* with given values $\mathbf{x}^\star$ is:

The $(1-\alpha)\times 100\%$*confidence interval* for the mean value of *y* with given values $\mathbf{x}^\star$ is:

Prediction

What can go wrong with predictions:

1. Poorly fitting model (overfitting, not enough data to fit properly).
2. Quantitative extrapolation.
3. Qualitative extrapolation.

Explanation

Explanation is another possible goal of regression. ***But what does explanation mean?***

1. In some data analysis problems, the parameters in a regression problem might have **physical meaning** (e.g., Hooke's Law). In those cases, explanation might be relatively clear (e.g., the spring constant in Hooke's Law characterizes the stiffness of the spring).

2. In other cases, it might be (relatively) easy to conduct an **experiment** to assign a causal interpretation to a given parameter.

3. But in many **observational studies**, parameters don't have physical meaning (e.g., meaning given by some theory). How do they help explain relationships?

Explanation

Causality and Experimental Design

The meaning of ***causality*** is an incredibly interesting area of study for philosophers, statisticians, computer scientists, etc. One definition of a causal effect is:

Definition*:* The ***causal effect*** of an action is the difference between the outcomes where the action was or was not taken.

Notice that there is a ***counterfactual*** involved in this definition: the causal effect of ***C*** involves knowing what ***would have happened if C did not occur.***

Explanation

Causality and Experimental Design: One way to try to estimate a causal effect is to design and conduct an experiment.

Example: Suppose that we want to compare the effects of high intensity interval training (HIIT) of low intensity steady-state training (LISS) on weight loss.

We could:

1. Collect a random sample (this may be hard!) from the population of interest.

2. ***Randomly*** split the sample into two groups: a HIIT group (*E =0*), a LISS group (*E = 1*).

3. Potentially measure other predictors (e.g., calorie intake).

The random assignment of exercise types assures that the two groups are not unbalanced in some way that favors either type of exercise. It will help us estimate the average difference in weight loss across these two groups, i.e., which type has a larger causal effect.

Explanation

Example: Suppose that we want to compare the effects of high intensity interval training (HIIT) of low intensity steady-state training (LISS) on weight loss.

$$Y_i = \beta_0 + \gamma E_i + \beta_1 X_{i,1} + \ldots + \beta_p X_{i,p} + \varepsilon_i$$

1. *Y* is weight loss.

2. E_i is the type of exercise assigned to the ith individual: 0 = HIIT, 1 = LISS

3. X_j are other variables that are known to influence weight (e.g., calorie intake).

Assuming that the design of the experiment is reasonable, we could interpret _____ as the average difference on weigh loss between HIIT and LISS, controlling for other measured variables.

Explanation

The design of experiments is very important, and we will return to some details later in the semester. For now, we'll say that, if an experiment is designed properly, then we can place a high degree of confidence in conclusions about causal effects.

But what about situations in which we don't (or can't!) have an experiment? In ***observational studies***, we might collect data about sample units without being able to randomly assign a "treatment".

Explanation

Example: Consider the following model for ***observational data***:

$$Y_i = \beta_0 + \gamma S_i + \beta_1 X_{i,1} + \ldots + \beta_p X_{i,p} + \varepsilon_i$$

where:

1. Y_i is wages per hour.

2. S_i is the sex of the ith individual*: 0 = male, 1 = female

3. X_j are other variables that are known to influence wages (e.g., years of education, geographic region of the country).

In this case, since S cannot be randomly assigned, the conclusions that we draw must be interpreted much more cautiously. Good data collection designs can certainly strengthen those conclusions. In any case, we might say, "controlling for many important factors, being female is associated with earning _____ less/more, on average."

Explanation

The literature on causation is enormous. If you're interested, check out some references:

1. An introduction to some philosophical issues in the metaphysics and epistemology of causation (https://plato.stanford.edu/entries/causation-metaphysics/)

2. A philosophical introduction to causal modeling (https://plato.stanford.edu/entries/causal-models/)

3. A new "general audience" book—*The Book of Why*—on causation and causal modeling by a pioneer in the field (https://www.amazon.com/Book-Why-Science-Cause-Effect/dp/046509760X)

4. A book chapter on causal inference using regression (http://www.stat.columbia.edu/~gelman/arm/chap9.pdf)