STAC67: Regression Analysis

Lecture 2

Thibault Randrianarisoa

September 11, 2026

University of Toronto Scarborough

Covariance and Correlation

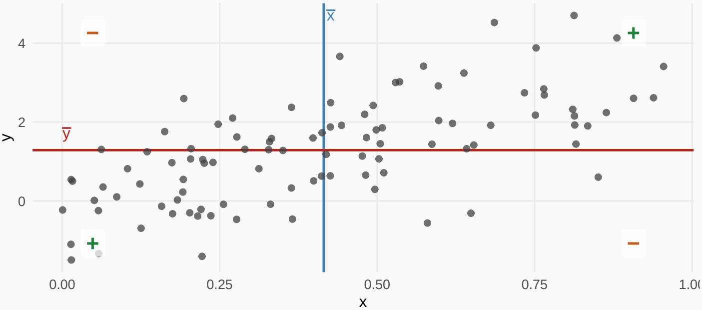
Covariance: $Cov(X, Y) = E((X - \mu_x)(Y - \mu_y))$, where $\mu_x = E(X),\ \mu_y = E(Y)$

For $Z_x = \frac{X - \mu_x}{\sqrt{Var(X)}},\ \ Z_y = \frac{Y - \mu_y}{\sqrt{Var(Y)}}$, the correlation coefficient is given by

$$Cov(Z_x, Z_y) = \rho_{xy} = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

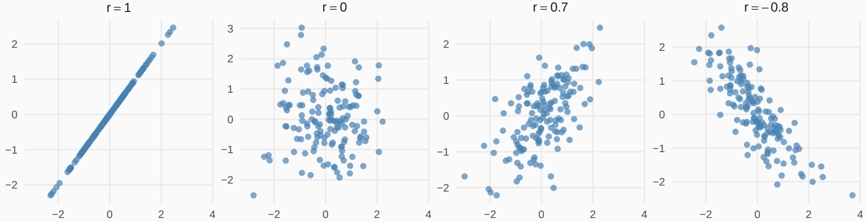
- $-1 \leq \rho_{xy} \leq 1$
- When the relationship is perfectly linear then $|\rho| = 1$.
- if two variables are independent then $\rho = 0$. (Note: the inverse does not hold)

Sample Covariance and Correlation



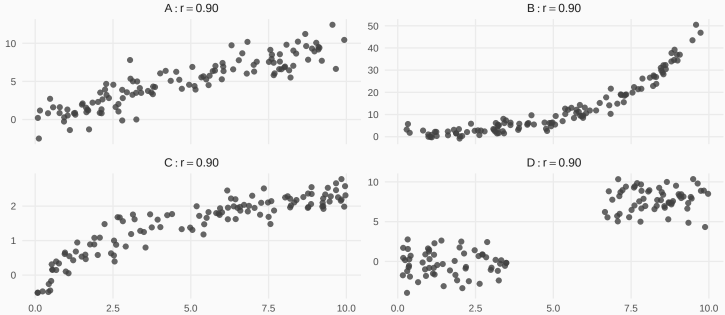
$$Cor(Y,X) = Cov(Z_y, Z_x) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_i - \bar{y}}{s_y}\right)\left(\frac{x_i - \bar{x}}{s_x}\right)$$

Correlation



Question: what are main differences between correlation and regression model?

Same r, different shapes



What is r actually measuring in each picture?

Test for Population correlation

When $\rho = 0$, and the joint distribution of (X, Y) is bivariate normal, and it can be shown that:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

has a Student's t distribution with $n-2$ degrees of freedom

$$H_0 : \rho = 0 \quad vs \quad H_1 : \rho \neq 0$$

Testing procedure

- Calculating the observed value of t (call this t_{obs})
- Compute the p-value for the test

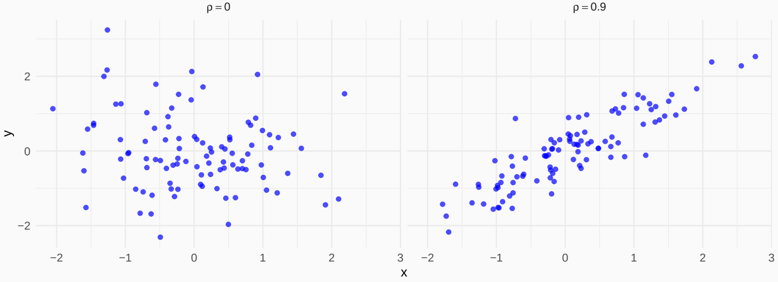
Simulation

```r
library(mvtnorm)
set.seed(123)
sigma.1 <- matrix(c(1, 0,   0,   1), ncol = 2)   # rho = 0
sigma.2 <- matrix(c(1, 0.9, 0.9, 1), ncol = 2)   # rho = 0.9

X1 <- rmvnorm(100, mean = c(0, 0), sigma.1)
X2 <- rmvnorm(100, mean = c(0, 0), sigma.2)

library(ggplot2)
sim <- rbind(data.frame(lab = "rho == 0",   x = X1[, 1], y = X1[, 2]),
             data.frame(lab = "rho == 0.9", x = X2[, 1], y = X2[, 2]))

ggplot(sim, aes(x, y)) +
  geom_point(size = 0.7, alpha = 0.7, colour = "blue") +
  facet_wrap(~ lab, labeller = label_parsed) +
  theme_minimal(base_size = 7)
```



Simulation

```r
x <- X2[, 1]
y <- X2[, 2]
cor.test(x, y)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 18.766, df = 98, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8327213 0.9209058
## sample estimates:
##       cor
## 0.8844736
```

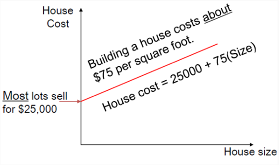
```r
r <- cor(x, y)
t <- r * sqrt(length(x) - 2) / sqrt(1 - r^2)
t
```

```
## [1] 18.76559
```

Relationship between variables

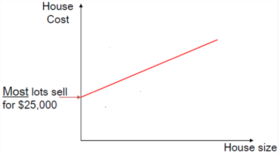
What factor or variable affects the price of house?

1) Mathematical relation: $Y = f(X)$, where X, Y are variables and f is a function



2) Statistical relation:

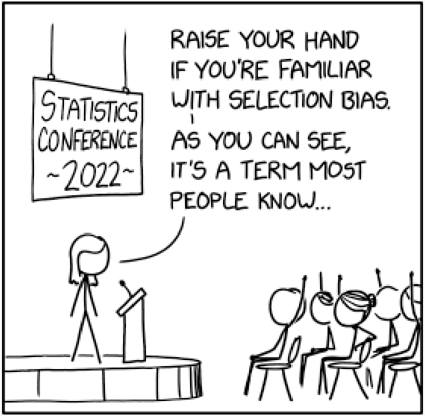
$$Y = f(X) + \epsilon$$



Data Collection for regression analysis

- Observational study
  - Investigator has no control over the explanatory variables (X)
  - Limitation: not adequate for cause-and-effect. A strong association does not necessarily mean a cause-and-effect relationship
- Experiment
  - Investigator exercises control over the explanatory variables (X) through active/random assignment
  - Random assignment balances out effect of other variables that might affect Y
  - Gold standard for cause-and-effect conclusions

Who ends up in your data?



Getting into the sample is itself an outcome and rarely a random one...

The Regression Process

1. The researcher must clearly define the question(s) of interest in the study
2. The response variable Y must be decided on, based on the question of interest.
3. A set of potentially relevant covariates, which can be measured, needs to be defined.
4. Data is collected.
5. Model Specification.
6. Decide on a method for fitting the specified model
7. Fit the model (typically using software such as R)
8. Examine the fitted model for violations of assumptions.
9. Conduct hypothesis testing for questions of interest.
10. Report the results from statistical inference.

Background Review: distributions

Normal distribution

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < \infty$$

Standard normal

$$Z = \frac{X-\mu}{\sigma} \sim N(0, 1^2)$$

Inference for a normal distribution. We have a random sample $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, with density $f(x; \theta)$ where θ is an unknown population parameter (a constant):

1. estimation of θ;
2. confidence interval for θ;
3. testing hypotheses about the value of θ.

In general,

- $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the unbiased estimator for μ
- $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is the unbiased estimator for σ^2

Background Review: sampling distributions

Sampling distribution of the estimators.

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) \qquad \boxed{T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)}$$

Exact when the population is normal; for any population with finite variance it holds approximately as $n \to \infty$, by the CLT.

Just check!

We have a random sample $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} N(\mu, 2^2)$.

1. What is the variance of the sample mean?
2. What is the variance of Y_{n+1}?
3. What is the standard error of $\bar{Y}$?

Background Review: sampling distributions

Sampling distribution of the estimators.

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) \qquad \boxed{T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)}$$

Exact when the population is normal; for any population with finite variance it holds approximately as $n \to \infty$, by the CLT.

Just check!

We have a random sample $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} N(\mu, 2^2)$.

1. What is the variance of the sample mean? $\sigma^2/n = 4/n$
2. What is the variance of Y_{n+1}?
3. What is the standard error of $\bar{Y}$?

Background Review: sampling distributions

Sampling distribution of the estimators.

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) \qquad \boxed{T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)}$$

Exact when the population is normal; for any population with finite variance it holds approximately as $n \to \infty$, by the CLT.

Just check!

We have a random sample $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} N(\mu, 2^2)$.

1. What is the variance of the sample mean? $\sigma^2/n = 4/n$
2. What is the variance of Y_{n+1}? $\sigma^2 = 4$
3. What is the standard error of $\bar{Y}$?

Background Review: sampling distributions

Sampling distribution of the estimators.

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) \qquad \boxed{T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)}$$

Exact when the population is normal; for any population with finite variance it holds approximately as $n \to \infty$, by the CLT.

Just check!

We have a random sample $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} N(\mu, 2^2)$.

1. What is the variance of the sample mean? $\sigma^2/n = 4/n$
2. What is the variance of Y_{n+1}? $\sigma^2 = 4$
3. What is the standard error of $\bar{Y}$? $\sigma/\sqrt{n} = 2/\sqrt{n}$

Simple Linear Regression

Suppose we have n observed pairs $(X_i, Y_i), i = 1, \ldots, n$.

Assumptions

1. A linear relationship

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

   - Y_i is the value of the response variable in the ith trial.
   - X_i is a known constant, namely, the value of the predictor variable in the ith trial.
   - β_0, β_1: regression model cofficients

     β_0: the intercept

     β_1: the slope

2. ϵ_i are random errors with zero mean, $E(\epsilon_i) = 0$, with common variance, $Var(\epsilon_i) = \sigma^2$, and pairwise independent.

Important Features

In the simple linear model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

1. The response variable Y, given X, is a sum of two terms:

- A constant term $\beta_0 + \beta_1 X$
- A random term ϵ

2. $E(Y_i) = \beta_0 + \beta_1 X_i$, where $E(Y_i)$ is a shortcut for $E(Y_i|X_i)$, the mean of Y when $X = X_i$

3. $Var(Y_i) = \sigma^2$, where $Var(Y_i)$ is a shortcut for $Var(Y_i|X_i)$, the variance of Y when $X = X_i$

4. The outcomes Y_i are pairwise independent because ϵ_i are pairwise independent.

Important Features

Example

Y = the monthly electricity bill, X = electricity used that month (kWh)

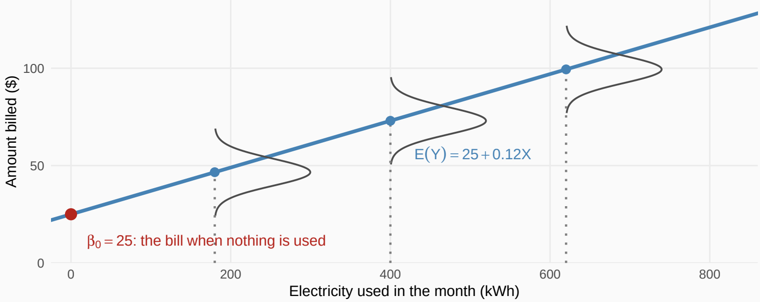
- Regression function: $E(Y) = 25 + 0.12\,X$
- $\beta_1 = 0.12$ indicates:
- $\beta_0 = 25$ indicates:

Important Features

Example

Y = the monthly electricity bill, X = electricity used that month (kWh)

- Regression function: $E(Y) = 25 + 0.12\,X$
- $\beta_1 = 0.12$ indicates:
- $\beta_0 = 25$ indicates:



Exercise

The regression model applies with $\beta_0 = 100, \ \ \beta_1 = 20$, and $\sigma^2 = 25$. An observation on Y will be made for $X = 5$.

a. Can you state the exact probability that Y will fall between 195 and 205? Explain.

b. If the normal error regression model is applicable, can you now state the exact probability that Y will fall between 195 and 205? If so, state it.

c. In the normal error regression model is applicable, a second observation on Y is made for $X = 6$. What is the exact probability that both observation fall between 195 and 205?