INTRODUCTION TO DATA SCIENCE

This lecture is based on course by
M. Cetinkaya-Rundel, Duke University, Data Analysis and Statistical Inference
and book by
M. Cetinkaya-Rundel and J. Handrin, „Introduction to Modern Statistics”

https://www.openintro.org/book/stat/

30/10 2025

WFAiS UJ, Informatyka Stosowana
I stopień studiów

Modern statistics

- **Scientists seek to answer questions using rigorous methods and careful observations**
- **These observations – collected from the likes of field notes, surveys, and experiments – form the blackbone of statistical investigations and are called data.**
- **Statistics is the study of how best collect, analyze and draw conclusions from data.**

Study design

Sampling principles and strategies

The first step in conducting research is to identify topics or questions that are to be investigated. A clearly laid out research question is helpful in identifying what subjects or cases should be studied and what variables are important. It is also important to consider *how* data are collected so that the data are reliable and help achieve the research goals.

A proficient analyst will have a good sense of the types of data they are working with and how to visualize the data in order to gain a complete understanding of the variables. Equally important, however, is the data source.

Sampling principles and strategies

Population and sample

Consider the following three research questions:

1. What is the average mercury content in swordfish in the Atlantic Ocean?
2. Over the last five years, what is the average time to complete a degree for Duke undergrads?
3. Does a new drug reduce the number of deaths in patients with severe heart disease?

Each research question refers to a target **population**. In the first question, the target population is all swordfish in the Atlantic Ocean, and each fish represents a case. Oftentimes, it is not feasible to collect data for every case in a population. Collecting data for an entire population is called a **census**. A census is difficult because it is too expensive to collect data for the entire population, but it might also be because it is difficult or impossible to identify the entire population of interest! Instead, a sample is taken. A **sample** is the data you have. Ideally, a sample is a small fraction of the population. For instance, 60 swordfish (or some other number) in the population might be selected, and this sample data may be used to provide an estimate of the population average and to answer the research question.

Experiments

Studies where the researchers assign treatments to cases are called **experiments**. When this assignment includes randomization, e.g., using a coin flip to decide which treatment a patient receives, it is called a **randomized experiment**. Randomized experiments are fundamentally important when trying to show a causal connection between two variables.

Principle of experimental design

1. **Controlling.** Researchers assign treatments to cases, and they do their best to **control** any other differences in the groups[3].

2. **Randomization.** Researchers randomize patients into treatment groups to account for variables that cannot be controlled.

3. **Replication.** The more cases researchers observe, the more accurately they can estimate the effect of the explanatory variable on the response. In a single study, we **replicate** by collecting a sufficiently large sample.

4. **Blocking.** Researchers sometimes know or suspect that variables, other than the treatment, influence the response. Under these circumstances, they may first group individuals based on this variable into **blocks** and then randomize cases within each block to the treatment groups. This strategy is often referred to as **blocking**.

Experiments

- **Blocking**

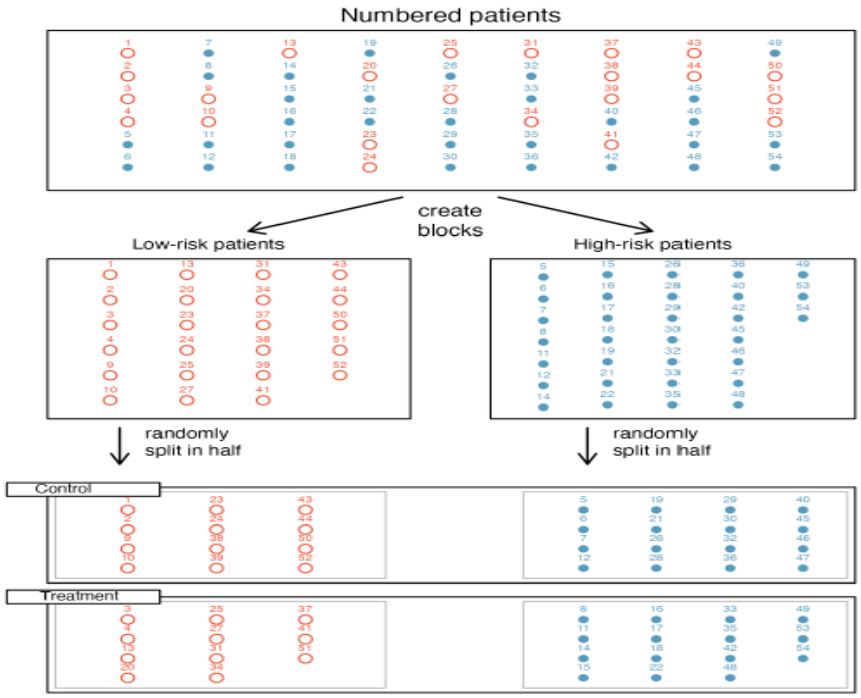


Figure 2.6: Blocking for patient risk. Patients are first divided into low-risk and high-risk blocks, then patients in each block are evenly randomized into the treatment groups. This strategy ensures equal representation of patients in each treatment group from both risk categories.

Experiments

Reducing bias in human experiments

Randomized experiments have long been considered to be the gold standard for data collection, but they do not ensure an unbiased perspective into the cause-and-effect relationship in all cases. Human studies are perfect examples where bias can unintentionally arise. Here we reconsider a study where a new drug was used to treat heart attack patients. In particular, researchers wanted to know if the drug reduced deaths in patients.

These researchers designed a randomized experiment because they wanted to draw causal conclusions about the drug's effect. Study volunteers[5] were randomly placed into two study groups. One group, the **treatment group**, received the drug. The other group, called the **control group**, did not receive any drug treatment.

Read more …. about placebo effect, blind study, double-blind setup

Observational studies

Studies where no treatment has been explicitly applied (or explicitly withheld) are called **observational studies**. For instance, studies on the loan data and county data described in Section 1.2 are would both be considered observational, as they rely on **observational data**.

Making causal conclusions based on experiments is often reasonable, since we can randomly assign the explanatory variable(s), i.e., the treatments. However, making the same causal conclusions based on observational data can be treacherous and is not recommended. Thus, observational studies are generally only sufficient to show associations or form hypotheses that can be later checked with experiments.

Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen *causes* skin cancer?

No! Some previous research tells us that using sunscreen actually reduces skin cancer risk, so maybe there is another variable that can explain this hypothetical association between sunscreen usage and skin cancer, as shown in Figure 2.7. One important piece of information that is absent is sun exposure. If someone is out in the sun all day, they are more likely to use sunscreen *and* more likely to get skin cancer. Exposure to the sun is unaccounted for in the simple observational investigation.

Observational studies

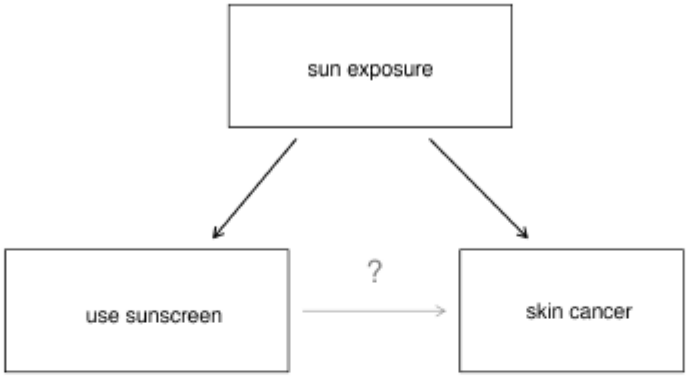
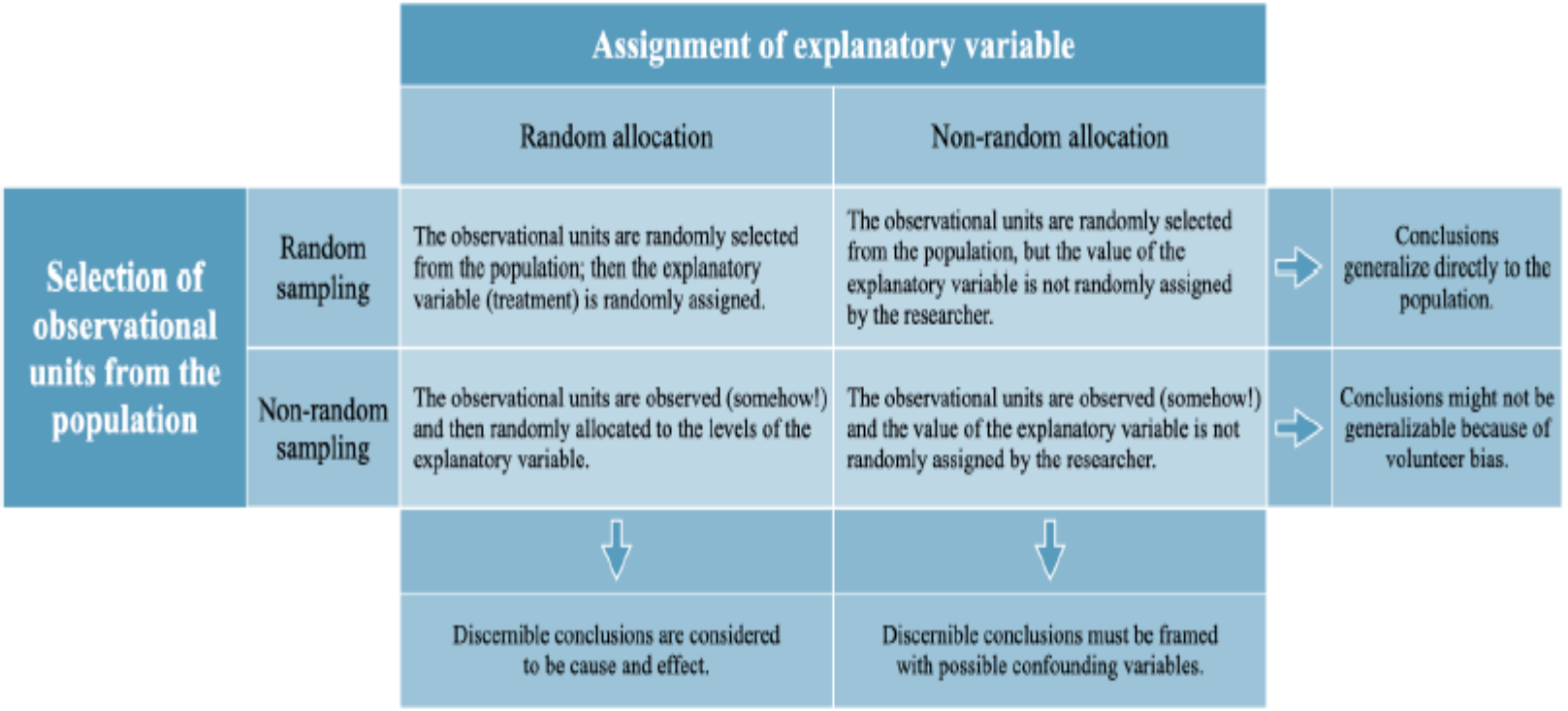


Figure 2.7: Sun exposure may be the root cause of both sunscreen use and skin cancer.

In this example, sun exposure is a confounding variable. The presence of confounding variables is what inhibits the ability for observational studies to make causal claims. While one method to justify making causal conclusions from observational studies is to exhaust the search for confounding variables, there is no guarantee that all confounding variables can be examined or measured.

Study design



		Assignment of explanatory variable		
		Random allocation	Non-random allocation	
Selection of observational units from the population	Random sampling	The observational units are randomly selected from the population; then the explanatory variable (treatment) is randomly assigned.	The observational units are randomly selected from the population, but the value of the explanatory variable is not randomly assigned by the researcher.	⇨ Conclusions generalize directly to the population.
	Non-random sampling	The observational units are observed (somehow!) and then randomly allocated to the levels of the explanatory variable.	The observational units are observed (somehow!) and the value of the explanatory variable is not randomly assigned by the researcher.	⇨ Conclusions might not be generalizable because of volunteer bias.
		⇩	⇩	
		Discernible conclusions are considered to be cause and effect.	Discernible conclusions must be framed with possible confounding variables.	

Figure 2.8: Analysis conclusions should be made carefully according to how the data were collected. Very few datasets come from the top left box because usually ethics require that random assignment of treatments can only be given to volunteers. Both representative (ideally random) sampling and experiments (random assignment of treatments) are important for how statistical conclusions can be made on populations.

Case study: Olympic 1500m

In this case study we introduce a dataset comparing Olympic and Paralympic gold medal finishers in the 1500m running competition (the Olympic "mile", if a bit shorter than a full mile).

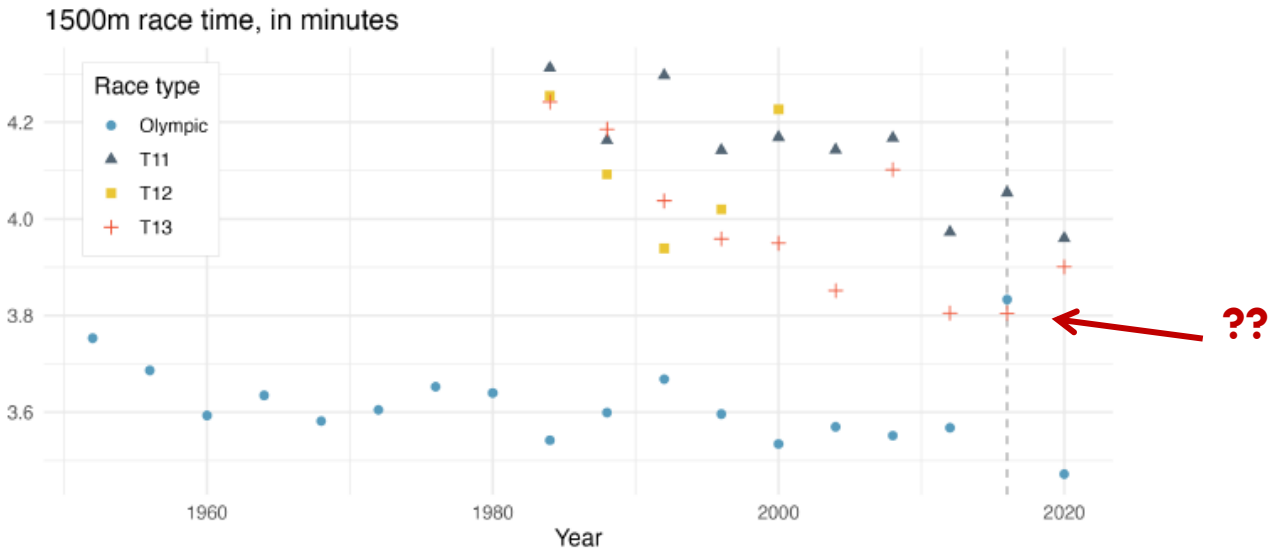


Figure 3.2: 1500m race time for Men's Olympic and Paralympic athletes. Dashed grey line represents the Rio Games in 2016.

The T11 athletes have almost complete visual impairment and are allowed to run with a guide-runner T12 and T13 athletes have some visual impairment

Case study: Olympic 1500m

Simson's paradox

Simpson's paradox is a description of three (or more) variables. The paradox happens when a third variable reverses the relationship between the first two variables.

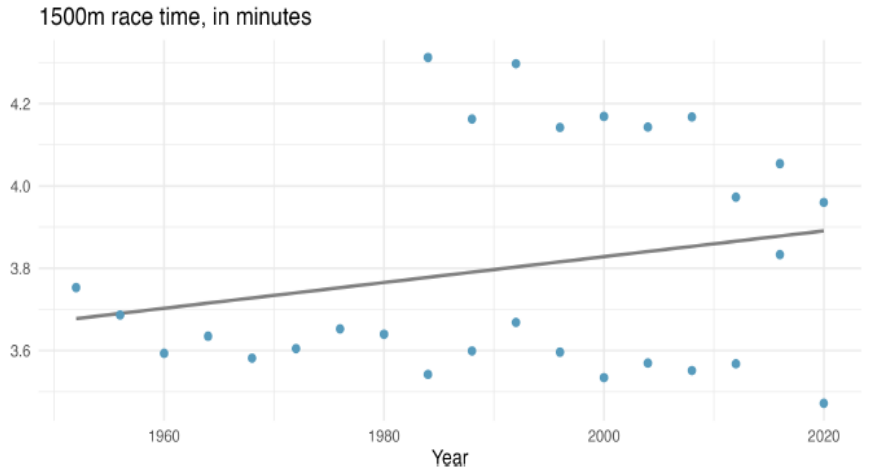


Figure 3.3: 1500m race time for Men's Olympic and Paralympic (T11) athletes. The line represents a line

Of course, both your eye and your intuition are likely telling you that it wouldn't make any sense to try to model all of the athletes together. Instead, a separate model should be run for each of the two types of Games: Olympic and Paralympic (T11).

Case study: Olympic 1500m

Simson's paradox

Simpson's paradox is a description of three (or more) variables. The paradox happens when a third variable reverses the relationship between the first two variables.

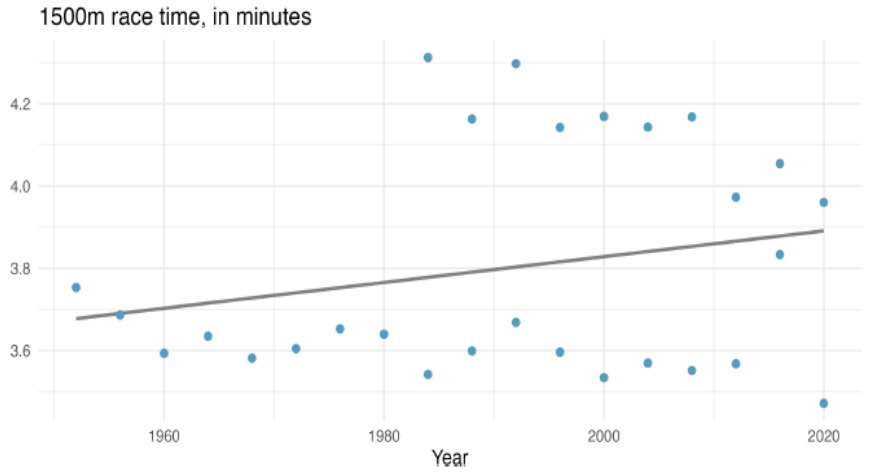


Figure 3.3: 1500m race time for Men's Olympic and Paralympic (T11) athletes. The line represents a line

Of course, both your eye and your intuition are likely telling you that it wouldn't make any sense to try to model all of the athletes together. Instead, a separate model should be run for each of the two types of Games: Olympic and Paralympic (T11).

Case study: Olympic 1500m

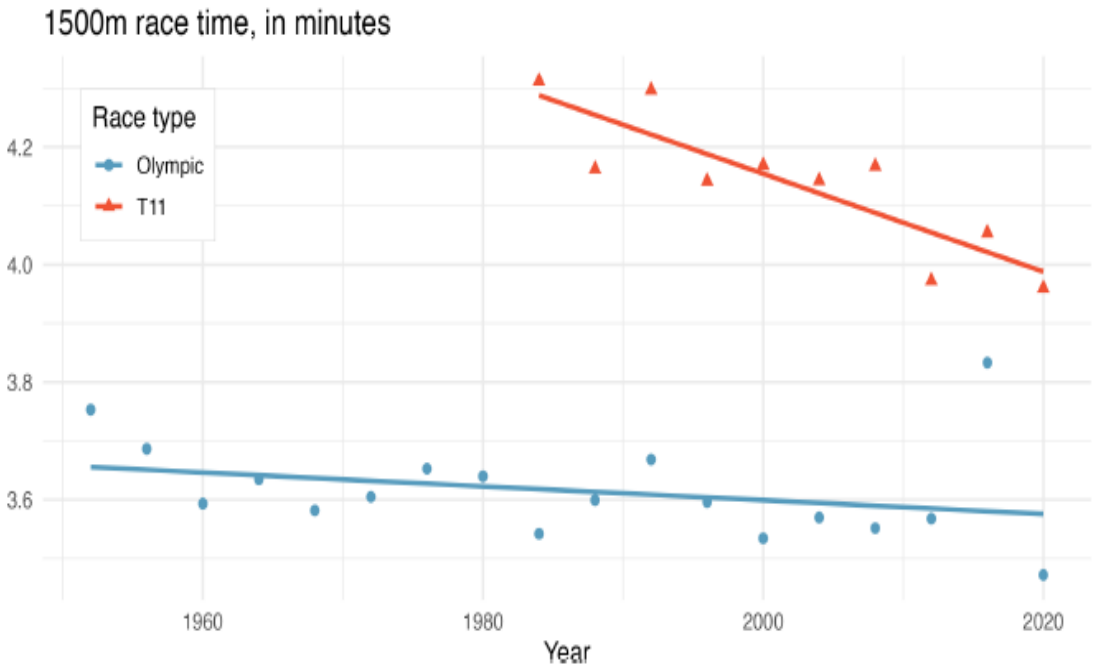


Figure 3.4: 1500m race time for Men's Olympic and Paralympic (T11) athletes. The best fit line is now fit separately to the Olympic and Paralympic athletes.

Case study: Olympic 1500m

Simpson's paradox.

Simpson's paradox happens when an association or relationship between two variables in one direction (e.g., positive) reverses (e.g., becomes negative) when a third variable is considered.

In the 1500m analysis, it would be most prudent to report the trends separately for the Olympic and the T11 athletes. However, in other situations, it might be better to aggregate the data and report the overall trend.

Regression modeling

- We skip most of it here, as covered during lectures in October; for more examples read chapters in the book
- Here included only what I found explained differently or not covered during lectures in October

Linear regression with single predictor

Linear regression is a very powerful statistical technique. Many people have some familiarity with regression models just from reading the news, where straight lines are overlaid on scatterplots. Linear models can be used for prediction or to describe the relationship between two numerical variables, assuming there is a linear relationship between them.

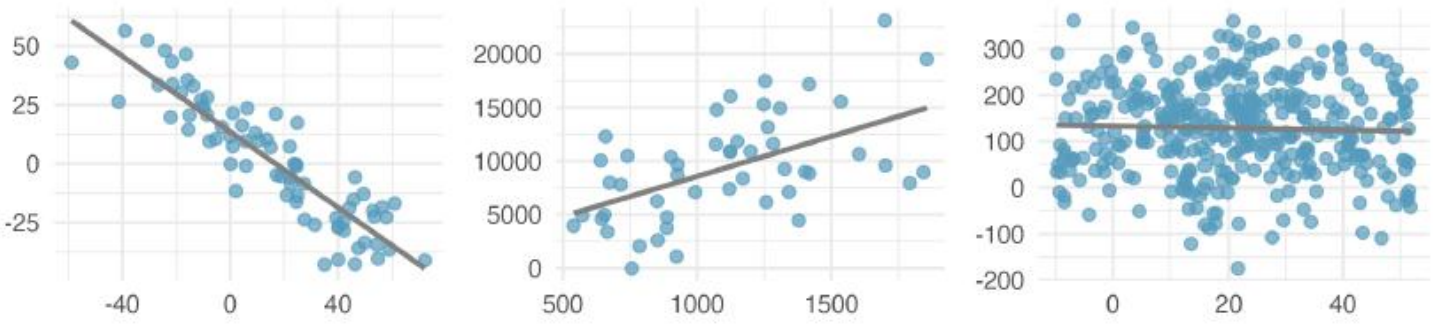


Figure 7.2: Three datasets where a linear model may be useful even though the data do not all fall exactly on the line.

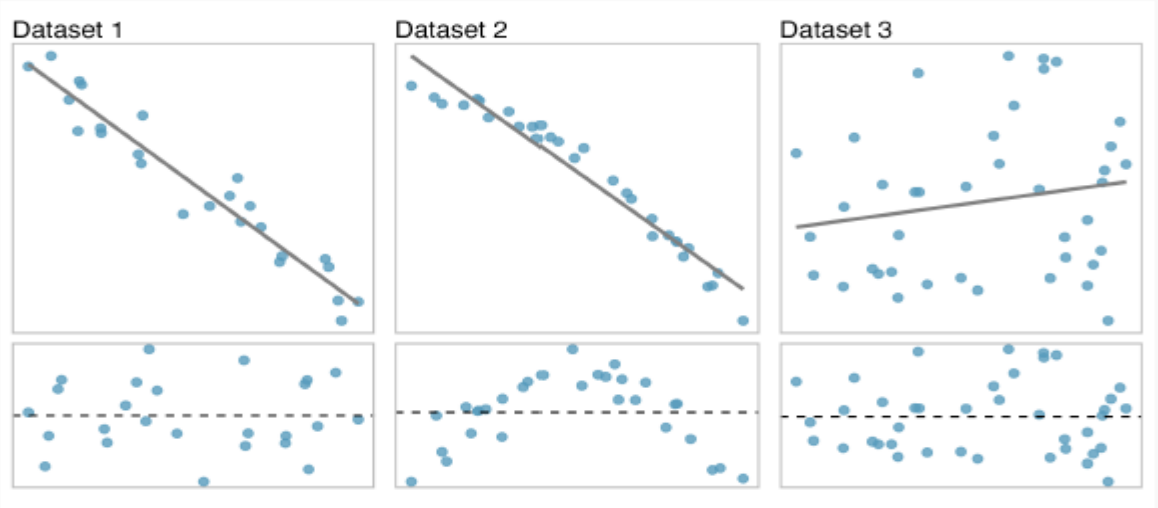
Residuals

Residuals are the leftover variation in the data after accounting for the model fit:

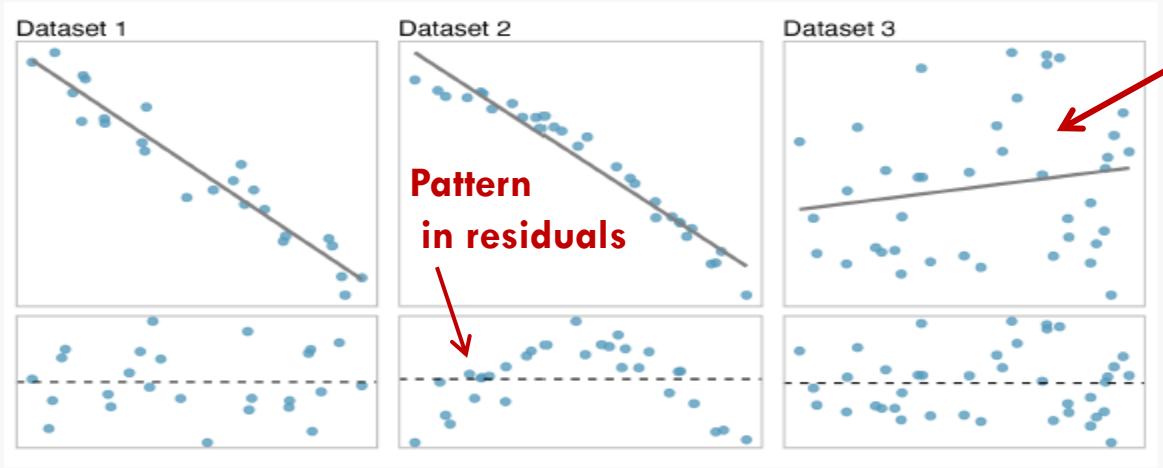
$$\text{Data} = \text{Fit} + \text{Residual}$$

EXAMPLE

One purpose of residual plots is to identify characteristics or patterns still apparent in data after fitting a model. The figure below shows three scatterplots with linear models in the first row and residual plots in the second row. Can you identify any patterns in the residuals?



Residuals



Dataset 1: the residuals show no obvious patterns. The residuals are scattered randomly around 0, represented by the dashed line.

Dataset 2: The second dataset shows a pattern in the residuals. There is some curvature in the scatterplot, which is more obvious in the residual plot. We should not use a straight line to model these data. Instead, a more advanced technique should be used to model the curved relationship, such as the variable transformations discussed in Section 5.7.

Dataset 3: The last plot shows very little upwards trend, and the residuals also show no obvious patterns. It is reasonable to try to fit a linear model to the data. However, it is unclear whether there is evidence that the slope parameter is different from zero. The point estimate of the slope parameter is not zero, but we might wonder if this could just be due to chance.

Describing linear relationship with correlations

Correlation: strength of a linear relationship.

Correlation which always takes values between -1 and 1, describes the strength and direction of the linear relationship between two variables. We denote the correlation by r.

The correlation value has no units and will not be affected by a linear change in the units (e.g., going from inches to centimeters).

We can compute the correlation using a formula

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$$

where $\bar{x}$, $\bar{y}$, s_x, and s_y are the sample means and standard deviations for each variable.

Describing linear relationship with correlations

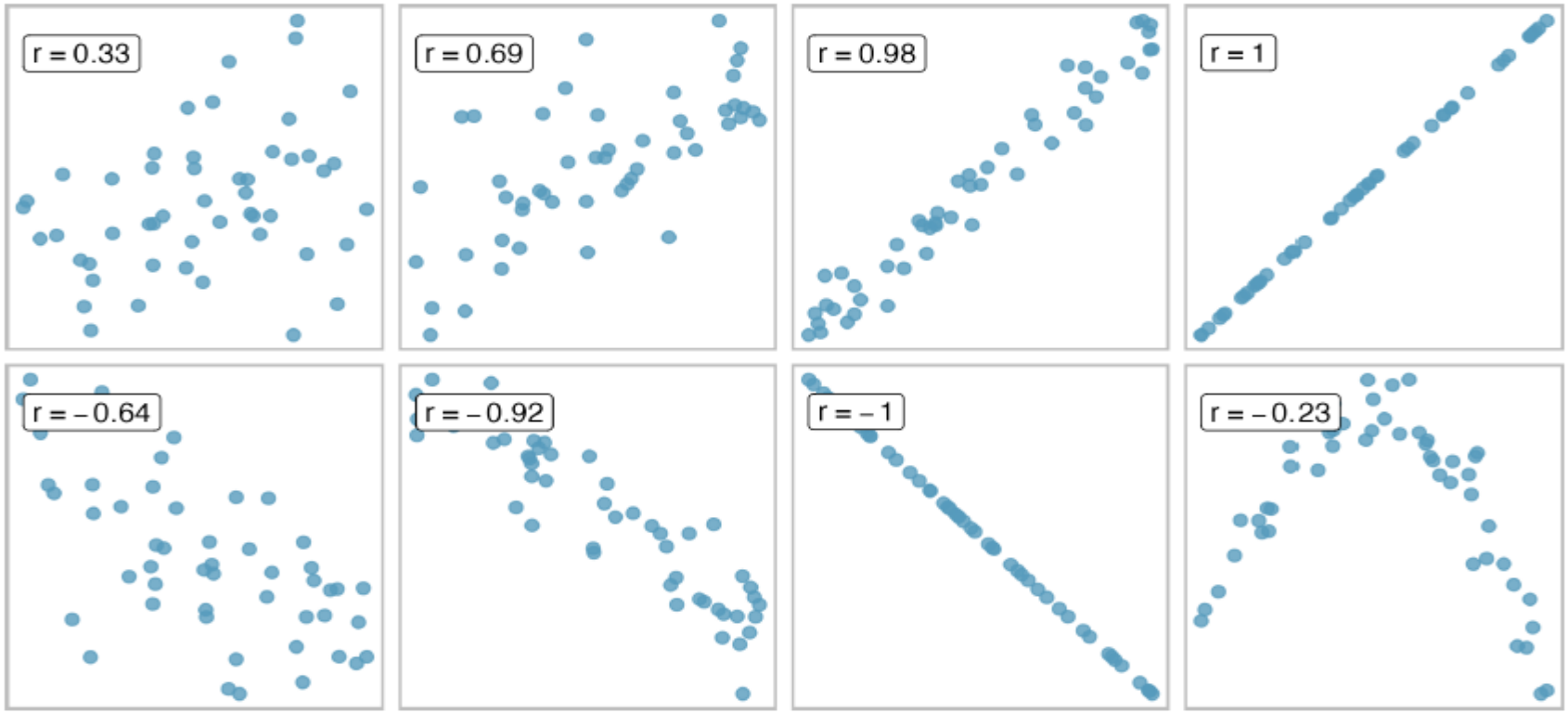


Figure 7.10: Sample scatterplots and their correlations. The first row shows variables with a positive relationship, represented by the trend up and to the right. The second row shows variables with a negative trend, where a large value in one variable is associated with a lower value in the other.

Describing linear relationship with correlations

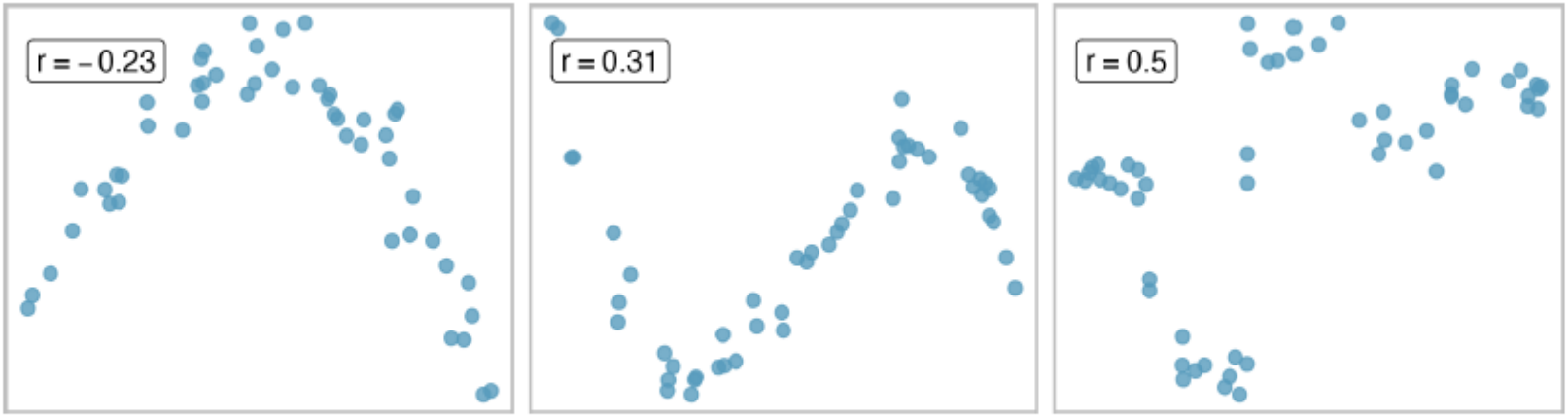


Figure 7.11: Sample scatterplots and their correlations. In each case, there is a strong relationship between the variables. However, because the relationship is not linear, the correlation is relatively weak.

Categorical predictors with two levels

Categorical variables are also useful in predicting outcomes. Here we consider a categorical predictor with two levels (recall that a *level* is the same as a *category*).

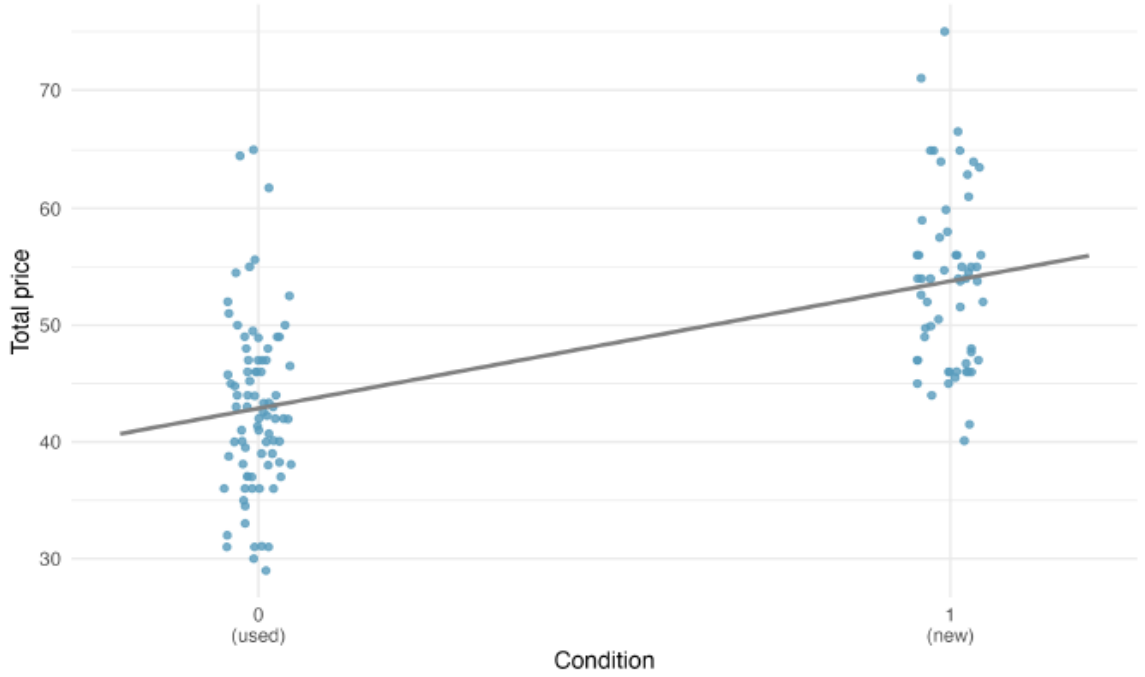


Figure 7.15: Total auction prices for the video game Mario Kart, divided into used ($x = 0$) and new ($x = 1$) condition games. The least squares regression line is also shown.

Categorical predictors with two levels

To incorporate the game condition variable into a regression equation, we must convert the categories into a numerical form. We will do so using an **indicator variable** called `condnew`, which takes value 1 when the game is new and 0 when the game is used. Using this indicator variable, the linear model may be written as

$$\widehat{\texttt{price}} = b_0 + b_1 \times \texttt{condnew}$$

The parameter estimates are given in Table 7.4.

Table 7.4: Least squares regression summary for the final auction price against the condition of the game.

term	estimate	std.error	statistic	p.value
(Intercept)	42.9	0.81	52.67	<0.0001
condnew	10.9	1.26	8.66	<0.0001

Using values from Table 7.4, the model equation can be summarized as

$$\widehat{\texttt{price}} = 42.87 + 10.9 \times \texttt{condnew}$$

Categorical predictors with two levels

Interpreting model estimates for categorical predictors.

The estimated intercept is the value of the outcome variable for the first category (i.e., the category corresponding to an indicator value of 0). The estimated slope is the average change in the outcome variable between the two categories.

The intercept is the estimated price when `condnew` has a value 0, i.e., when the game is in used condition. That is, the average selling price of a used version of the game is \$42.9. The slope indicates that, on average, new games sell for about \$10.9 more than used games.

Note that, fundamentally, the intercept and slope interpretations do not change when modeling categorical variables with two levels. However, when the predictor variable is binary, the coefficient estimates (b_0 and b_1) are directly interpretable with respect to the dataset at hand.

Coefficient of determination: R-squared

Sums of squares to measure variability in y.

We can measure the variability in the y values by how far they tend to fall from their mean, $\bar{y}$. We define this value as the total sum of squares, calculated using the formula below, where y_i represents each y value in the sample, and $\bar{y}$ represents the mean of the y values in the sample.

$$SST = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2.$$

Left-over variability in the y values if we know x can be measured by the sum of squared errors, or sum of squared residuals, calculated using the formula below, where $\hat{y}_i$ represents the predicted value of y_i based on the least squares regression.[8],

$$\begin{aligned} SSE &= (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2 \\ &= e_1^2 + e_2^2 + \cdots + e_n^2 \end{aligned}$$

The coefficient of determination can then be calculated as

$$R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

[8]The difference $SST - SSE$ is called the regression sum of squares, SSR, and can also be calculated as $SSR = (\hat{y}_1 - \bar{y})^2 + (\hat{y}_2 - \bar{y})^2 + \cdots + (\hat{y}_n - \bar{y})^2$. SSR represents the variation in y that was accounted for in our model.

Outliers

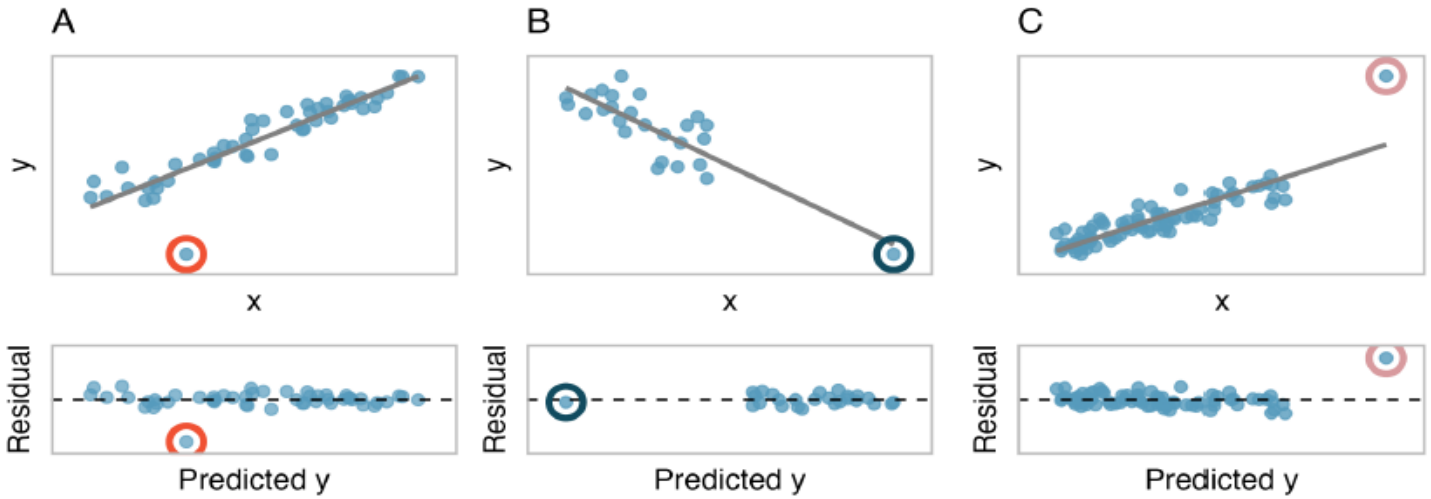
Types of outliers.

A point (or a group of points) that stands out from the rest of the data is called an outlier. Outliers that fall horizontally away from the center of the cloud of points are called leverage points. Outliers that influence on the slope of the line are called influential points.

Leverage.

Points that fall horizontally away from the center of the cloud tend to pull harder on the line, so we call them points with high leverage or leverage points.

Outliers

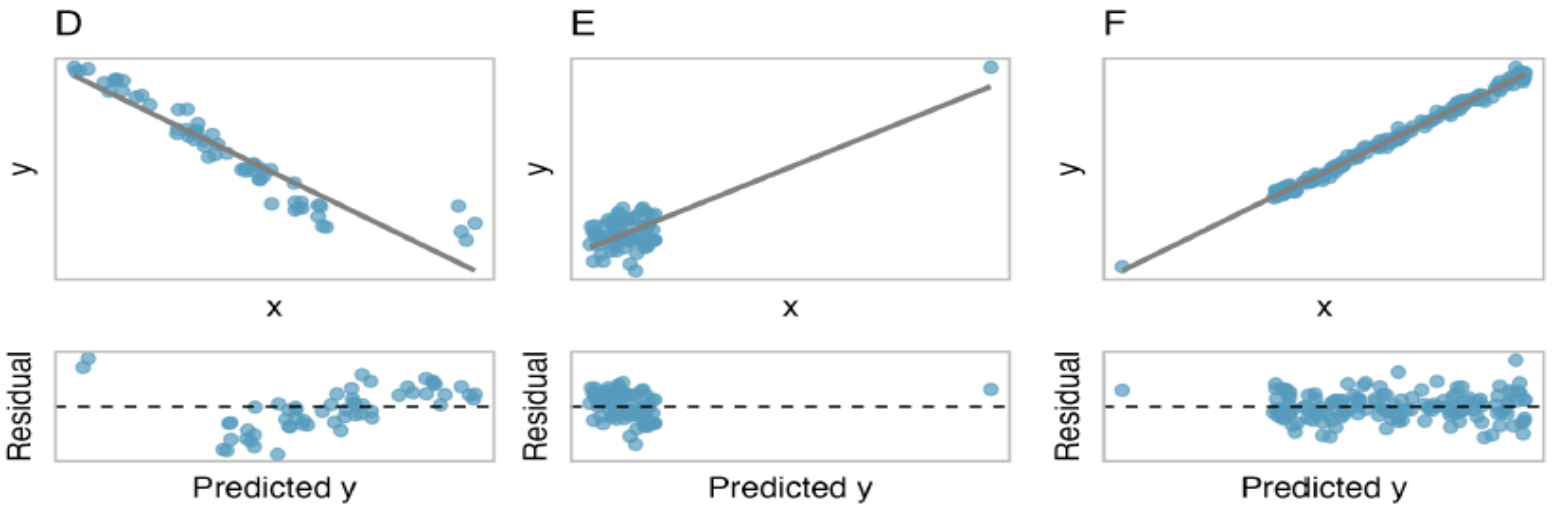


A: There is one outlier far from the other points (in the y direction and it is an outlier of the bivariate model), though it only appears to slightly influence the line.

B: There is one outlier on the right (in the x and y direction although it is not an outlier of the bivariate model), though it is quite close to the least squares line, which suggests it wasn't very influential.

C: There is one point far away from the cloud (in the x and y direction and an outlier of the bivariate model), and this outlier appears to pull the least squares line up on the right; examine how the line around the primary cloud does not appear to fit very well.

Outliers



D: There is a primary cloud and then a small secondary cloud of four outliers (with respect to both x and the bivariate model). The secondary cloud appears to be influencing the line somewhat strongly, making the least square line fit poorly almost everywhere. There might be an interesting explanation for the dual clouds, which is something that could be investigated.

E: There is no obvious trend in the main cloud of points and the outlier on the right (with respect to both x and y) appears to largely (and problematically) control the slope of the least squares line. The point creates a bivariate model when seemingly there is none.

F: There is one outlier far from the cloud (with respect to both x and y). However, it falls quite close to the least squares line and does not appear to be very influential (it is not outlying with respect to the bivariate model).

Outliers

A good practice for dealing with outlying observations is to produce two analyses: one with and one without the outlying observations. Presenting both analyses to a client and discussing the role of the outlying observations should lead you to a more holistic understanding of the appropriate model for the data.

Points that fall horizontally far from the line are points of high leverage; these points can strongly influence the slope of the least squares line. If one of these high leverage points does appear to actually invoke its influence on the slope of the line – as in Plots C, D, and E of Figure 7.16a and Figure 7.16b – then we call it an **influential point**. Usually we can say a point is influential if, had we fitted the line without it, the influential point would have been unusually far from the least squares line.

It is tempting to remove outliers. Don't do this without a very good reason. Models that ignore exceptional (and interesting) cases often perform poorly. For instance, if a financial firm ignored the largest market swings – the "outliers" – they would soon go bankrupt by making poorly thought-out investments.

Linear regression with multiple predictors

Multiple regression extends single predictor variable regression to the case that still has one response but many predictors (denoted x_1, x_2, x_3, ...). The method is motivated by scenarios where many variables may be simultaneously connected to an output.

Building on the ideas of one predictor variable in a linear regression model (from Chapter 7), a multiple linear regression model is now fit to two or more predictor variables. By considering how different explanatory variables interact, we can uncover complicated relationships between the predictor variables and the response variable. One challenge to working with multiple variables is that it is sometimes difficult to know which variables are most important to include in the model. Model building is an extensive topic, and we scratch the surface here by defining and utilizing the adjusted R^2 value.

Linear regression with multiple predictors

The dataset includes information on 10000 loans

Table 8.1: First six rows of the **loans** dataset.

interest_rate	verified_income	debt_to_income	credit_util	bankruptcy	term	credit_checks	issue_month
14.07	Verified	18.01	0.548	0	60	6	Mar-2018
12.61	Not Verified	5.04	0.150	1	36	1	Feb-2018
17.09	Source Verified	21.15	0.661	0	36	4	Feb-2018
6.72	Not Verified	10.16	0.197	0	36	0	Jan-2018
14.07	Verified	57.96	0.755	0	36	7	Mar-2018
6.72	Not Verified	6.46	0.093	0	36	6	Jan-2018

Variable	Description
`interest_rate`	Interest rate on the loan, in an annual percentage.
`verified_income`	Categorical variable describing whether the borrower's income source and amount have been verified, with levels Verified (source and amount verified), Source Verified (source only verified), and Not Verified.
`debt_to_income`	Debt-to-income ratio, which is the percentage of total debt of the borrower divided by their total income.
`credit_util`	Of all the credit available to the borrower, what fraction are they utilizing. For example, the credit utilization on a credit card would be the card's balance divided by the card's credit limit.
`bankruptcy`	An indicator variable for whether the borrower has a past bankruptcy in their record. This variable takes a value of '1' if the answer is *yes* and '0' if the answer is *no*.
`term`	The length of the loan, in months.
`issue_month`	The month and year the loan was issued, which for these loans is always during the first quarter of 2018.
`credit_checks`	Number of credit checks in the last 12 months. For example, when filing an application for a credit card, it is common for the company receiving the application to run a credit check.

Linear regression with multiple predictors

Indicator and categorical predictors

Let's start by fitting a linear regression model for interest rate with a single predictor indicating whether a person has a bankruptcy in their record:

$$\widehat{\texttt{interest\_rate}} = 12.34 + 0.74 \times \texttt{bankruptcy}$$

Table 8.4: Summary of a linear model for predicting `interest_rate` based on whether the borrower has a bankruptcy in their record. Degrees of freedom for this model is 9998.

term	estimate	std.error	statistic	p.value
(Intercept)	12.34	0.05	231.49	<0.0001
bankruptcy1	0.74	0.15	4.82	<0.0001

The variable takes one of two values: 1 when the borrower has a bankruptcy in their history and 0 otherwise. A slope of 0.74 means that the model predicts a 0.74% higher interest rate for those borrowers with a bankruptcy in their record.

Linear regression with multiple predictors

Indicator and categorical predictors

Suppose we had fit a model using a 3-level categorical variable, such as `verified_income`. The output from software is shown in Table 8.5. This regression output provides multiple rows for the variable. Each row represents the relative difference for each level of `verified_income`. However, we are missing one of the levels: `Not Verified`. The missing level is called the **reference level** and it represents the default level that other levels are measured against.

Table 8.5: Summary of a linear model for predicting `interest_rate` from the borrower's income source and amount verification. This predictor has three levels, which results in 2 rows in the regression output.

term	estimate	std.error	statistic	p.value
`(Intercept)`	11.10	0.08	137.2	<0.0001
`verified_incomeSource Verified`	1.42	0.11	12.8	<0.0001
`verified_incomeVerified`	3.25	0.13	25.1	<0.0001

Linear regression with multiple predictors

Table 8.5: Summary of a linear model for predicting `interest_rate` from the borrower's income source and amount verification. This predictor has three levels, which results in 2 rows in the regression output.

term	estimate	std.error	statistic	p.value
(Intercept)	11.10	0.08	137.2	<0.0001
verified_incomeSource Verified	1.42	0.11	12.8	<0.0001
verified_incomeVerified	3.25	0.13	25.1	<0.0001

EXAMPLE

How would we write an equation for this regression model?

The equation for the regression model may be written as a model with two predictors:

$$\widehat{\texttt{interest\_rate}} = 11.10 + 1.42 \times \texttt{verified\_income}_{\texttt{Source Verified}} + 3.25 \times \texttt{verified\_income}_{\texttt{Verified}}$$

We use the notation $\texttt{variable}_{\texttt{level}}$ to represent indicator variables for when the categorical variable takes a particular value. For example, $\texttt{verified\_income}_{\texttt{Source Verified}}$ would take a value of 1 if it was for a borrower that was source verified, and it would take a value of 0 otherwise. Likewise, $\texttt{verified\_income}_{\texttt{Verified}}$ would take a value of 1 if it was for a borrower that was verified, and 0 if it took any other value.

Linear regression with multiple predictors

When `verified_income` takes a value of `Not Verified`, then both indicator functions in the equation for the linear model are set to 0:

$$\widehat{\texttt{interest\_rate}} = 11.10 + 1.42 \times 0 + 3.25 \times 0 = 11.10$$

The average interest rate for these borrowers is 11.1%. Because the level does not have its own coefficient and it is the reference value, the indicators for the other levels for this variable all drop out.

When `verified_income` takes a value of `Source Verified`, then the corresponding variable takes a value of 1 while the other is 0:

$$\widehat{\texttt{interest\_rate}} = 11.10 + 1.42 \times 1 + 3.25 \times 0 = 12.52$$

The average interest rate for these borrowers is 12.52%.

Linear regression with multiple predictors

The higher interest rate for borrowers who have verified their income source or amount is surprising. Intuitively, we would think that a loan would look *less* risky if the borrower's income has been verified. However, note that the situation may be more complex, and there may be confounding variables that we didn't account for. For example, perhaps lenders require borrowers with poor credit to verify their income. That is, verifying income in our dataset might be a signal of some concerns about the borrower rather than a reassurance that the borrower will pay back the loan. For this reason, the borrower could be deemed higher risk, resulting in a higher interest rate.

[1]When `verified_income` takes a value of `Verified`, then the corresponding variable takes a value of 1 while the other is 0: $11.10 + 1.42 \times 0 + 3.25 \times 1 = 14.35$. The average interest rate for these borrowers is 14.35%.

[2]Each of the coefficients gives the incremental interest rate for the corresponding level relative to the `Not Verified` level, which is the reference level. For example, for a borrower whose income source and amount have been verified, the model predicts that they will have a 3.25% higher interest rate than a borrower who has not had their income source or amount verified.

[3]Relative to the `Not Verified` category, the `Verified` category has an interest rate of 3.25% higher, while the `Source Verified` category is only 1.42% higher. Thus, `Verified` borrowers will tend to get an interest rate about 3.25 higher than `Source Verified` borrowers.

Linear regression with multiple predictors

Predictors with several categories.

Software = R package

When fitting a regression model with a categorical variable that has k levels where $k > 2$, software will provide a coefficient for $k - 1$ of those levels. For the last level that does not receive a coefficient, this is the reference level, and the coefficients listed for the other levels are all considered relative to this reference level.

The higher interest rate for borrowers who have verified their income source or amount is surprising. Intuitively, we would think that a loan would look *less* risky if the borrower's income has been verified. However, note that the situation may be more complex, and there may be confounding variables that we didn't account for. For example, perhaps lenders require borrowers with poor credit to verify their income. That is, verifying income in our dataset might be a signal of some concerns about the borrower rather than a reassurance that the borrower will pay back the loan. For this reason, the borrower could be deemed higher risk, resulting in a higher interest rate.

Many predictors in the model

The world is complex, and it can be helpful to consider many factors at once in statistical modeling.

We want to construct a model that accounts not only for any past bankruptcy or whether the borrower had their income source or amount verified, but simultaneously accounts for all the variables in the `loans` dataset: `verified_income`, `debt_to_income`, `credit_util`, `bankruptcy`, `term`, `issue_month`, and `credit_checks`.

$$\begin{aligned}\widehat{\texttt{interest\_rate}} = b_0 &\\ &+ b_1 \times \texttt{verified\_income}_{\texttt{Source Verified}} + b_2 \times \texttt{verified\_income}_{\texttt{Verified}} \\ &+ b_3 \times \texttt{debt\_to\_income} + b_4 \times \texttt{credit\_util} \\ &+ b_5 \times \texttt{bankruptcy} + b_6 \times \texttt{term} \\ &+ b_7 \times \texttt{credit\_checks} + b_8 \times \texttt{issue\_month}_{\texttt{Jan-2018}} \\ &+ b_9 \times \texttt{issue\_month}_{\texttt{Mar-2018}}\end{aligned}$$

This equation represents a holistic approach for modeling all of the variables simultaneously. Notice that there are two coefficients for `verified_income` and two coefficients for `issue_month`, since both are 3-level categorical variables.

Many predictors in the model

The world is complex, and it can be helpful to consider many factors at once in statistical modeling.

The fitted model for the interest rate is given by:

$$\begin{aligned}\widehat{\texttt{interest\_rate}} = 1.89 \\ &+ 1.00 \times \texttt{verified\_income}_{\texttt{Source Verified}} + 2.56 \times \texttt{verified\_income}_{\texttt{Verified}} \\ &+ 0.02 \times \texttt{debt\_to\_income} + 4.90 \times \texttt{credit\_util} \\ &+ 0.39 \times \texttt{bankruptcy} + 0.15 \times \texttt{term} \\ &+ 0.23 \times \texttt{credit\_checks} + 0.05 \times \texttt{issue\_month}_{\texttt{Jan-2018}} \\ &- 0.04 \times \texttt{issue\_month}_{\texttt{Mar-2018}}\end{aligned}$$

If we count up the number of predictor coefficients, we get the *effective* number of predictors in the model; there are nine of those. Notice that the categorical predictor counts as two, once for each of the two levels shown in the model. In general, a categorical predictor with p different levels will be represented by $p - 1$ terms in a multiple regression model. A total of seven variables were used as predictors to fit this model: `verified_income`, `debt_to_income`, `credit_util`, `bankruptcy`, `term`, `credit_checks`, `issue_month`.

Adjusted R-squared

Adjusted R-squared as a tool for model assessment.

The adjusted R-squared is computed as

$$R^2_{adj} = 1 - \frac{s^2_{\text{residuals}}/(n-k-1)}{s^2_{\text{outcome}}/(n-1)} = 1 - \frac{s^2_{\text{residuals}}}{s^2_{\text{outcome}}} \times \frac{n-1}{n-k-1}$$

where n is the number of observations used to fit the model and k is the number of predictor variables in the model. Remember that a categorical predictor with p levels will contribute $p-1$ to the number of variables in the model.

Stepwise selection using adjusted R^2 as the decision criteria is one of many commonly used model selection strategies. Stepwise selection can also be carried out with decision criteria other than adjusted R^2, such as p-values, AIC (Akaike information criterion) or BIC (Bayesian information criterion)

Alternatively, one could choose to include or exclude predictors from a model based on expert opinion or due to research focus. In fact, many statisticians discourage the use of stepwise regression *alone* for model selection and advocate, instead, for a more thoughtful approach that carefully considers the research focus and features of the data.

Logistic regression

In this chapter we introduce **logistic regression** as a tool for building models when there is a categorical response variable with two levels, e.g., yes and no. Logistic regression is a type of **generalized linear model (GLM)** for response variables where regular multiple regression does not work very well. GLMs can be thought of as a two-stage modeling approach. We first model the response variable using a probability distribution, such as the binomial or Poisson distribution. Second, we model the parameter of the distribution using a collection of predictors and a special form of multiple regression. Ultimately, the application of a GLM will feel very similar to multiple regression, even if some of the details are different.

Logistic regression: discrimination of hiring case

We will consider experiment data from a study that sought to understand the effect of race and sex on job application callback rates (Bertrand and Mullainathan 2003).

Table 9.1: List of all 36 unique names along with the commonly inferred race and sex associated with these names.

first_name	race	sex	first_name	race	sex	first_name	race	sex
Aisha	Black	female	Hakim	Black	male	Laurie	White	female
Allison	White	female	Jamal	Black	male	Leroy	Black	male
Anne	White	female	Jay	White	male	Matthew	White	male
Brad	White	male	Jermaine	Black	male	Meredith	White	female
Brendan	White	male	Jill	White	female	Neil	White	male
Brett	White	male	Kareem	Black	male	Rasheed	Black	male
Carrie	White	female	Keisha	Black	female	Sarah	White	female
Darnell	Black	male	Kenya	Black	female	Tamika	Black	female
Ebony	Black	female	Kristen	White	female	Tanisha	Black	female
Emily	White	female	Lakisha	Black	female	Todd	White	male
Geoffrey	White	male	Latonya	Black	female	Tremayne	Black	male
Greg	White	male	Latoya	Black	female	Tyrone	Black	male

Race and sex are protected classes in the United States, meaning they are not legally permitted factors for hiring or employment decisions.

The response variable of interest is whether there was a callback from the employer for the applicant

Logistic regression: discrimination of hiring case

Table 9.2: Descriptions of nine variables from the `resume` dataset. Many of the variables are indicator variables, meaning they take the value 1 if the specified characteristic is present and 0 otherwise.

variable	description
`received_callback`	Specifies whether the employer called the applicant following submission of the application for the job.
`job_city`	City where the job was located: Boston or Chicago.
`college_degree`	An indicator for whether the resume listed a college degree.
`years_experience`	Number of years of experience listed on the resume.
`honors`	Indicator for the resume listing some sort of honors, e.g. employee of the month.
`military`	Indicator for if the resume listed any military experience.
`has_email_address`	Indicator for if the resume listed an email address for the applicant.
`race`	Race of the applicant, implied by their first name listed on the resume.
`sex`	Sex of the applicant (limited to only man and woman), implied by the first name listed on the resume.

All of the attributes listed on each resume were randomly assigned, which means that no attributes that might be favorable or detrimental to employment would favor one demographic over another on these resumes. Importantly, due to the experimental nature of the study, we can infer causation between these variables and the callback rate, if substantial differences are found. Our analysis will allow us to compare the practical importance of each of the variables relative to each other.

Logistic regression

Logistic regression is a generalized linear model where the outcome is a two-level categorical variable. The outcome, Y_i, takes the value 1 (in our application, the outcome represents a callback for the resume) with probability p_i and the value 0 with probability $1 - p_i$. Because each observation has a slightly different context, e.g., different education level or a different number of years of experience, the probability p_i will differ for each observation. Ultimately, it is the **probability** of the outcome taking the value 1 (i.e., being a "success") that we model in relation to the predictor variables: we will examine which resume characteristics correspond to higher or lower callback rates.

Notation for a logistic regression model.

The outcome variable for a GLM is denoted by Y_i, where the index i is used to represent observation i. In the resume application, Y_i will be used to represent whether resume i received a callback ($Y_i = 1$) or not ($Y_i = 0$).

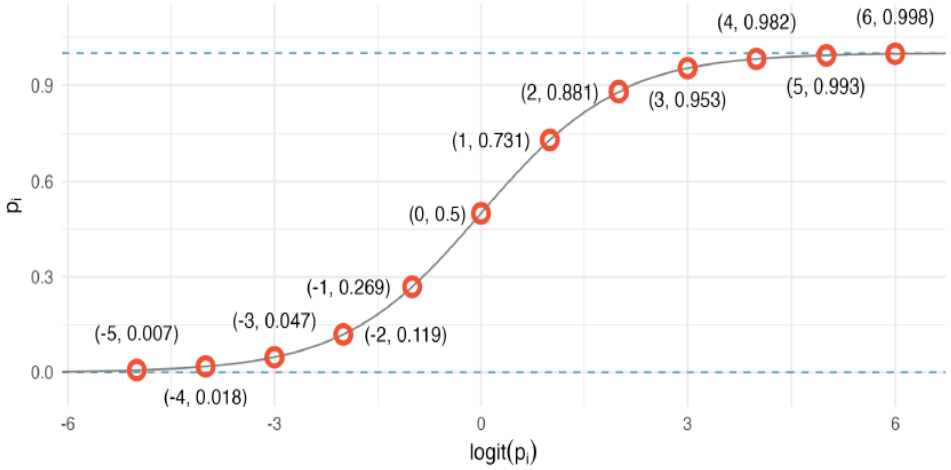
$$transformation(p_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i}$$

Logistic regression

We want to choose a **transformation** in the equation that makes practical and mathematical sense. A common transformation for p_i is the **logit transformation**, which may be written as

$$logit(p_i) = \log_e\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i}$$

The **logit transformation** is shown in Figure 9.1. Below, we rewrite the equation relating Y_i to its predictors using the logit transformation of p_i:



Modeling the probability of an event

We want to choose a **transformation** in the equation that makes practical and mathematical sense.

To convert from values on the logistic regression scale to the probability scale, we need to back transform and then solve for p_i:

$$\log_e\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}$$

$$\frac{p_i}{1-p_i} = e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}$$

$$p_i = (1-p_i)\, e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}$$

$$p_i = e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}} - p_i \times e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}$$

$$p_i + p_i\, e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}} = e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}$$

$$p_i(1 + e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}) = e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}$$

$$p_i = \frac{e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}}{1 + e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}}$$

As with most applied data problems, we substitute in the point estimates (the observed b_i) to calculate relevant probabilities.

Modeling the probability of an event

We start by fitting a model with a single predictor: **honors**. This variable indicates whether the applicant had any type of honors listed on their resume, such as employee of the month. A logistic regression model was fit using statistical software and the following model was found:

$$\log_e\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -2.4998 + 0.8668 \times \textbf{honors}$$

a. If a resume is randomly selected from the study and it does not have any honors listed, what is the probability it resulted in a callback?

b. What would the probability be if the resume did list some honors?

---

a. If a randomly chosen resume from those sent out is considered, and it does not list honors, then **honors** takes the value of 0 and the right side of the model equation equals -2.4998. Solving for p_i: $\frac{e^{-2.4998}}{1+e^{-2.4998}} = 0.076$. Just as we labeled a fitted value of y_i with a "hat" in single-variable and multiple regression, we do the same for this probability: $\hat{p}_i = 0.076$.

b. If the resume had listed some honors, then the right side of the model equation is $-2.4998 + 0.8668 \times 1 = -1.6330$, which corresponds to a probability $\hat{p}_i = 0.163$. Notice that we could examine -2.4998 and -1.6330 in Figure 9.1 to estimate the probability before formally calculating the value.

Modeling the probability of an event

Table 9.6: Summary table for the logistic regression model for the resume callback example, where vari selection has been performed using AIC and `college_degree` has been dropped from the model.

term	estimate	std.error	statistic	p.value
(Intercept)	-2.72	0.16	-17.51	<0.0001
job_cityChicago	-0.44	0.11	-3.83	1e-04
years_experience	0.02	0.01	2.02	0.043
honors1	0.76	0.19	4.12	<0.0001
military1	-0.34	0.22	-1.60	0.1105
has_email_address1	0.22	0.11	1.97	0.0494
raceWhite	0.44	0.11	4.10	<0.0001
sexman	-0.20	0.14	-1.45	0.1473

The `race` variable had taken only two levels: `Black` and `White`. Based on the model results, what does the coefficient of the `race` variable say about callback decisions?

The coefficient shown corresponds to the level of `White`, and it is positive. The positive coefficient reflects a positive gain in callback rate for resumes where the candidate's first name implied they were White. The model results suggest that prospective employers favor resumes where the first name is typically interpreted to be White.

Modeling the probability of an event

EXAMPLE

Use the model summarized in Table 9.6 to estimate the probability of receiving a callback for a job in Chicago where the candidate lists 14 years experience, no honors, no military experience, includes an email address, and has a first name that implies they are a White male.

We can start by writing out the equation using the coefficients from the model:

$$log_e\left(\frac{\hat{p}}{1-\hat{p}}\right) = -2.7162 - 0.4364 \times \texttt{job\_city}_{\texttt{Chicago}} + 0.0206 \times \texttt{years\_experience} + 0.7634 \times \texttt{honors} - 0.3443 \times \texttt{military} + 0.2221 \times \texttt{email} + 0.4429 \times \texttt{race}_{\texttt{White}} - 0.1959 \times \texttt{sex}_{\texttt{man}}$$

Now we can add in the corresponding values of each variable for the individual of interest:

$$log_e\left(\frac{\hat{p}}{1-\hat{p}}\right) = -2.7162 - 0.4364 \times 1 + 0.0206 \times 14 + 0.7634 \times 0 - 0.3443 \times 0 + 0.2221 \times 1 + 0.4429 \times 1 - 0.1959 \times 1 = -2.3955$$

We can now back-solve for $\hat{p}$: the chance such an individual will receive a callback is about $\frac{e^{-2.3955}}{1+e^{-2.3955}} = 0.0835$.