Beyond simple linear regression

- **Multiple regression:** more than one predictor/explanatory variable

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n,$$

- **Logistic regression:** predicting categorical outcomes with two possible outcomes.

Multiple regression

Multiple regression model

A multiple regression model is a linear model with many predictors. In general, we write the model as

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

when there are k predictors.

- Reading the tables
- Interpretation of the intercept and the coefficients
- R^2 for the multidimensional case
- Model selection; which predictors should we use in our model.
- Checking model assumption

Mario Kart

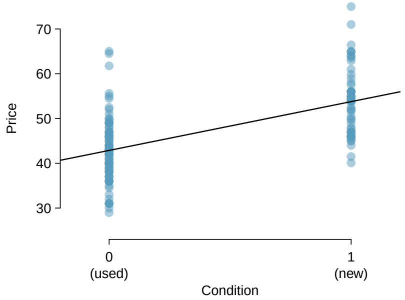
	price	cond_new	stock_photo	duration	wheels
1	51.55	1	1	3	1
2	37.04	0	1	7	1
⋮	⋮	⋮	⋮	⋮	⋮
140	38.76	0	0	7	0
141	54.51	1	1	1	2

variable	**description**
price	final auction price plus shipping costs, in US dollars
cond_new	a coded two-level categorical variable, which takes value 1 when the game is new and 0if the game is used
stock_photo	a coded two-level categorical variable, which takes value 1 if the primary photo used in the auction was a stock photo and 0 if the photo was unique to that auction
duration	the length of the auction, in days, taking values from 1 to 10
wheels	the number of Wii wheels included with the auction (a *Wii wheel* is a plastic racing wheel that holds the Wii controller and is an optional but helpful accessory for playing Mario Kart)

Simple linear regression

$$\widehat{price} = 42.87 + 10.90 \times cond\_new$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	42.8711	0.8140	52.67	0.0000
cond_new	10.8996	1.2583	8.66	0.0000
				$df = 139$



Adding more predictors

Multiple regression

Include all the potentially important variables simultaneously.

$$\begin{aligned}\widehat{\text{price}} &= \beta_0 + \beta_1 \times \text{cond\_new} + \beta_2 \times \text{stock\_ photo} \\ &\quad + \beta_3 \times \text{duration} + \beta_4 \times \text{wheels}\end{aligned}$$

Or

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Fitting the model

Estimate the parameters β_0, β_1, ..., β_4.
Process(similar to simple regression):

- Evaluate the sum squared residuals

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

- Select b_0, b_1, ..., b_4 that minimise SSE.

Fitting the model

In practice we retrieve the following table.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.2110	1.5140	23.92	0.0000
cond_new	5.1306	1.0511	4.88	0.0000
stock_photo	1.0803	1.0568	1.02	0.3085
duration	-0.0268	0.1904	-0.14	0.8882
wheels	7.2852	0.5547	13.13	0.0000
				$df = 136$

The multiple regression model is:

$$\hat{y} = 36.21 + 5.13x_1 + 1.08x_2 - 0.027x_3 + 7.29x_4$$

Interpretation of the parameters

$$\hat{y} = 36.21 + 5.13x_1 + 1.08x_2 - 0.027x_3 + 7.29x_4$$

- The estimated coefficient of variable x_4 (Wii wheels) is $\beta_4 = 7.29$.
  This implies that the average difference in auction price for each additional Wii wheel included, **holding all the other variables constant**, is 7.29.

- The estimated intercept is 36.21, this is the model's predicted price when each of the variables take value zero. However, when the auction duration is 0, that implies that the auction has not started yet, so the price must be zero. Hence, the intercept does not provide any insight in this case.

Adjusted R^2

- Simple linear regression

$$R^2 = 1 - \frac{\text{variability in residuals}}{\text{variability in the outcome}} = 1 - \frac{Var(e)}{Var(y)}$$

- Adjusted R^2 for multiple regression

$$R^2_{adj} = 1 - \frac{Var(e)/(n-k-1)}{Var(y)/(n-1)} = 1 - \frac{Var(e)}{Var(y)} \times \frac{n-1}{n-k-1}$$

where n is the number of cases used to fit the model and k is the number of predictor variables in the model.

Model Selection

- A model that includes all available explanatory variables is often referred to as the **full model.**
- Is the full model the optimal model?
  - Not necessarily.
- Explore different model selection strategies.

Mario Kart (again!)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.2110	1.5140	23.92	0.0000
cond_new	5.1306	1.0511	4.88	0.0000
stock_photo	1.0803	1.0568	1.02	0.3085
duration	-0.0268	0.1904	-0.14	0.8882
wheels	7.2852	0.5547	13.13	0.0000
$R^2_{adj} = 0.7108$				$df = 136$

The last column of the table lists p-values that can be used to assess hypotheses of the following form:

H_0: $\beta_i = 0$ when the other explanatory variables are included in the model.

H_A: $\beta_i \neq 0$ when the other explanatory variables are included in the model.

Selecting predictors

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.2110	1.5140	23.92	0.0000
cond_new	5.1306	1.0511	4.88	0.0000
stock_photo	1.0803	1.0568	1.02	0.3085
duration	-0.0268	0.1904	-0.14	0.8882
wheels	7.2852	0.5547	13.13	0.0000
$R^2_{adj} = 0.7108$				$df = 136$

Identify the variables in the model that may not be helpful.

Selecting predictors

- The $p-$value for the auction duration is 0.8882, which indicates that there is not statistically significant evidence that the duration is related to the total auction price when accounting for the other variables.
- The $p-$value for the condition of the game is zero, which indicates there is strong evidence that a game's condition (new or used) has a real relationship with the total auction price, when accounting for the other variables..

Two model selection strategies

Stepwise model selection strategies

- Backward selection
- Forward selection

Model selection strategy

Backward elimination strategy

- Start from a **full** model
- Drop the variable with the largest $p-$value
- Refit the model
- Drop the variable with the largest $p-$value
- Refit the model...

Repeat until happy...
Note: In the case, two variables have the same $p-$value examine the adjusted R^2.

Backward selection on Mario

The multiple regression model is:

$$\hat{y} = 36.21 + 5.13x_1 + 1.08x_2 - 0.027x_3 + 7.29x_4$$

with

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.2110	1.5140	23.92	0.0000
cond_new	5.1306	1.0511	4.88	0.0000
stock_photo	1.0803	1.0568	1.02	0.3085
duration	-0.0268	0.1904	-0.14	0.8882
wheels	7.2852	0.5547	13.13	0.0000
$R^2_{adj} = 0.7108$				$df = 136$

Which variable should we eliminate?

Backwards selection on Mario

After eliminating the **duration** variable, we refit the model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.0483	0.9745	36.99	0.0000
cond_new	5.1763	0.9961	5.20	0.0000
stock_photo	1.1177	1.0192	1.10	0.2747
wheels	7.2984	0.5448	13.40	0.0000
$R^2_{adj} = 0.7128$				$df = 137$

Which variable should we eliminate?

Backwards selection on Mario

After eliminating the **stock photo** variable, we refit the model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.7849	0.7066	52.06	0.0000
cond_new	5.5848	0.9245	6.04	0.0000
wheels	7.2328	0.5419	13.35	0.0000
$R^2_{adj} = 0.7124$				$df = 138$

Since all the $p-$values are equal to zero, we conclude that this is the optimal model given the variables.

Model selection strategy

Forward elimination strategy

- Start from a model that includes no variables
- Fit each of the possible models with just one variable.
- Select the variable with the smallest $p-$values.
- Add the variable to the model
- Expand this model by adding one of the remaining variables.
- Fit the respective models
- Select the variable with the smallest $p-$values.
- Add the variable to the model...

Repeat until happy...

Note: In the case, two variables have the same $p-$value examine the adjusted R^2.

Summary

- The **backward-elimination** strategy begins with the largest model and eliminates variables one-by-one until we are satisfied that all remaining variables are important to the model.
- The **forward-selection** strategy starts with no variables included in the model, then it adds in variables according to their importance until no other important variables are found.

Model assumptions

Are the model's assumptions satisfied?

Assumptions

Multiple regression methods using the model

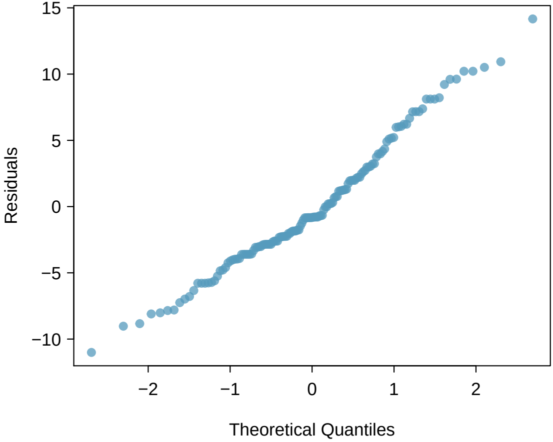
$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

generally depend on the following four **assumptions**:

1. the residuals of the model are nearly normal,
2. the variability of the residuals is nearly constant,
3. the residuals are independent, and
4. each variable is linearly related to the outcome.

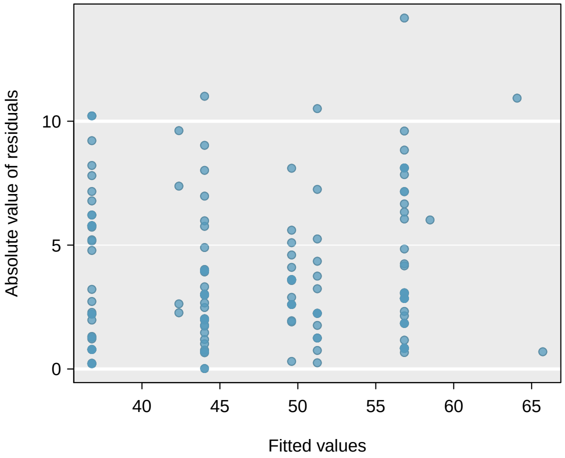
Nearly normal residuals

Normal probability plot for the residuals



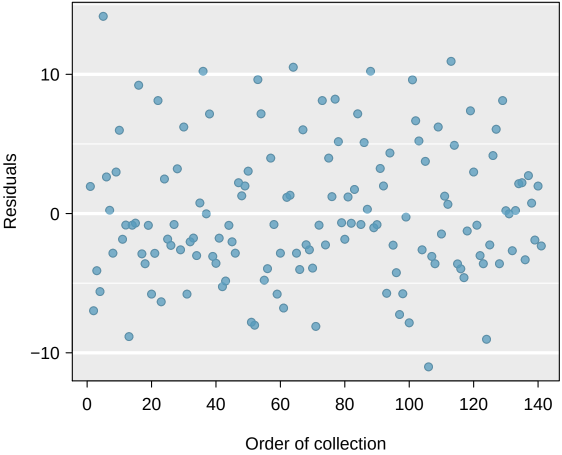
Near constant variability

Absolute values of residuals against fitted values



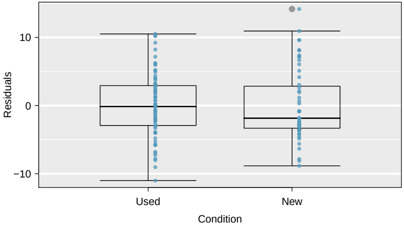
Residuals are independent

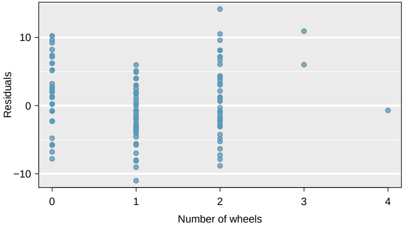
Residuals in order of their data collection



Each variable is linearly related to the outcome

Residuals against each predictor variable





We should not see any trend in the residual plots.

Notes

Matrix representation of data

Multi-regression model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

If we have n observations then,

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

where x_{ij} is the value of the *jth* independent variable in the *ith* observation, $j = 1, \ldots, n$.

Notes

Matrix notation

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{11} & x_{12} & \dots & x_{1k} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \qquad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Hence, the matrix representation of the multi-regression model is

$$Y = X\beta + \epsilon$$

Notes

Transformation of categorical variables

- **Two levels** categorical variable, ex. F/M, Yes/No, can be represented as

$$x = \begin{cases} 0, & \text{if F} \\ 1, & \text{if M} \end{cases}$$

- **k-levels** categorical variable, ex. agree, disagree, na. Introduce $k-1$ dummy variables

$$x_1 = \begin{cases} 1, & \text{if agree} \\ 0, & \text{otherwise} \end{cases} \quad x_2 = \begin{cases} 1, & \text{if disagree} \\ 0, & \text{otherwise} \end{cases}$$

-

	x_1	x_2
agree	1	0
disagree	0	1
na	0	0

Logistic regression

For y a **numerical response** variable and k predictors, $x_1, x_2, \ldots, x_k$, our goal is to find a relation between them:

$$y = f(x_1, x_2, \ldots, x_k)$$

- For a **numerical variable** y, we can try a **multi-regression** model:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- What about a **categorical variable** y?

Spam again...

	spam	to_multiple	cc	attach	dollar	winner	inherit	password	format	re_subj	exclaim_subj
1	0	0	0	0	0	0	0	0	1	1	0
2	0	0	0	0	0	0	0	0	1	0	0
3	1	0	1	1	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	1	1	0	0
6	0	0	0	0	0	0	0	0	0	0	0

variable	**description**
spam	Specifies whether the message was spam.
to_multiple	An indicator variable for if more than one person was listed in the *To* field of the email.
cc	An indicator for if someone was CCed on the email.
attach	An indicator for if there was an attachment, such as a document or image.
dollar	An indicator for if the word "dollar" or dollar symbol ($) appeared in the email.
winner	An indicator for if the word "winner" appeared in the email message.
inherit	An indicator for if the word "inherit" (or a variation, like "inheritance") appeared in the email.
password	An indicator for if the word "password" was present in the email.
format	Indicates if the email contained special formatting, such as bolding, tables, or links
re_subj	Indicates whether "Re:" was included at the start of the email subject.
exclaim_subj	Indicates whether any exclamation point was included in the email subject.

Logistic regression

Explanatory variable:
multiple recipients, cc, dollar, winner: 0
inherit, password, format, re_subj, exclaim_subj:1
Response variable:
Spam:0 **or** 1?
Logistic regression: The probability that the email is spam, given this explanatory variables.

Logistic regression

- y : **categorical response variable** , with two levels $\{0, 1\}$
- $x_1, x_2, \ldots, x_k$: k **predictors.**

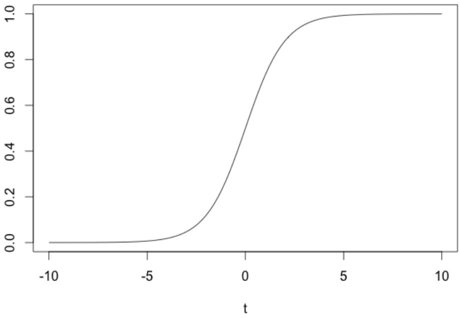
Then model for **logistic regression** is

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}},$$

where p is the probability of $y = 1$.

Logistic function

Logistic/Sigmoid function



$$f(t) = \frac{1}{1+e^{-t}}$$

Logistic regression

Likelihood function

$$L(y_1, \dots, y_n; x_1, \dots, x_k, \beta_0, \dots, \beta_k) = \\ \Pi_{i=1}^{n} p(x_{i1}, \dots, x_{ik})^{y_i} (1 - p(x_{i1}, \dots, x_{in}))^{1-y_i},$$

where $p(x_{i1}, \dots, x_{in}) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{i1}+\dots+\beta_k x_{ik})}}$.

The parameters for the logistic regression are the **maximum likelihood estimator**; $\beta_0, \dots, \beta_k$ that maximise $L(y_1, \dots, y_n; x_1, \dots, x_k, \beta_0, \dots, \beta_k)$.

Logistic regression

Logistic model for single predictor to_multiple.

$$p = \frac{1}{1 + e^{2.12 + 1.81x}}$$

If an email has no multiple recipients, to probability of it being spam is

$$p = \frac{1}{1 + e^{2.12}} = 0.11,$$

else the probability is

$$p = \frac{1}{1 + e^{2.12 + 1.81x}} = 0.02$$

Logistic regression

Logistic regression model for 11 predictors:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.8362	0.0962	-8.69	0.0000
to_multiple	-2.8836	0.3121	-9.24	0.0000
winner	1.7038	0.3254	5.24	0.0000
format	-1.5902	0.1239	-12.84	0.0000
re_subj	-2.9082	0.3708	-7.84	0.0000
exclaim_subj	0.1355	0.2268	0.60	0.5503
cc	-0.4863	0.3054	-1.59	0.1113
attach	0.9790	0.2170	4.51	0.0000
dollar	-0.0582	0.1589	-0.37	0.7144
inherit	0.2093	0.3197	0.65	0.5127
password	-1.4929	0.5295	-2.82	0.0048

Logistic regression

After variable selection.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.8595	0.0910	-9.44	0.0000
to_multiple	-2.8372	0.3092	-9.18	0.0000
winner	1.7370	0.3218	5.40	0.0000
format	-1.5569	0.1207	-12.90	0.0000
re_subj	-3.0482	0.3630	-8.40	0.0000
attach	0.8643	0.2042	4.23	0.0000
password	-1.4871	0.5290	-2.81	0.0049

Logistic regression

$$p = \frac{1}{1 + e^{-0.859 - 2.837x_1 + 1.737x_2 - 1.557x_3 - 3.05x_4 + 0.854x_5 - 1.487x_6}}$$