Regression analysis using Python

Eric Marsden
<eric.marsden@risk-engineering.org>

Some individuals use statistics as the drunken man uses lamp posts: for support rather than for illumination.
– attributed to Andrew Lang
Regression analysis

▷ Linear regression analysis means “fitting a straight line to data”
  • also called *linear modelling*

▷ It’s a widely used technique to help **model** and **understand**
  real-world phenomena
  • easy to use
  • easy to understand intuitively

▷ Allows **prediction**
A regression problem is composed of

- an outcome or response variable $Y$
- a number of risk factors or predictor variables $X_i$ that affect $Y$
  - also called explanatory variables, or features in the machine learning community
- a question about $Y$, such as *How to predict $Y$ under different conditions?*

$Y$ is sometimes called the *dependent variable* and $X_i$ the *independent variables*

- not the same meaning as *statistical independence*
- experimental setting where the $X_i$ variables can be modified and changes in $Y$ can be observed
Regression analysis: objectives

**Prediction**

We want to estimate $Y$ at some specific values of $X_i$.

**Model inference**

We want to learn about the relationship between $Y$ and $X_i$, such as the combination of predictor variables which has the most effect on $Y$. 
Univariate linear regression

(when all you have is a single predictor variable)
Linear regression

- Linear regression: one of the simplest and most commonly used statistical modeling techniques

- Makes strong assumptions about the relationship between the predictor variables \((X_i)\) and the response \((Y)\)
  - (a linear relationship, a straight line when plotted)
  - only valid for *continuous* outcome variables (not applicable to category outcomes such as *success/failure*)

\[
y = \beta_0 + \beta_1 \times x + \text{error}
\]
Linear regression

▷ Assumption: \( y = \beta_0 + \beta_1 \times x + \text{error} \)

▷ Our task: estimate \( \beta_0 \) and \( \beta_1 \) based on the available data

▷ Resulting model is \( \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times x \)
  • the “hats” on the variables represent the fact that they are estimated from the available data
  • \( \hat{y} \) is read as “the estimator for \( y \)”

▷ \( \beta_0 \) and \( \beta_1 \) are called the model parameters or coefficients

▷ Objective: minimize the error, the difference between our observations and the predictions made by our linear model
  • minimize the length of the red lines in the figure to the right (called the “residuals”)
Ordinary Least Squares regression

- Ordinary Least-Squares (OLS) regression: a method for selecting the model parameters
  - $\beta_0$ and $\beta_1$ are chosen to minimize the square of the distance between the predicted values and the actual values
  - equivalent to minimizing the size of the red rectangles in the figure to the right

- An application of a quadratic loss function
  - in statistics and optimization theory, a loss function, or cost function, maps from an observation or event to a number that represents some form of “cost”
Simple linear regression: example

- The *British Doctors’ Study* followed the health of a large number of physicians in the UK over the period 1951–2001.

- Provided conclusive evidence of linkage between smoking and lung cancer, myocardial infarction, respiratory disease and other illnesses.

- Provides data on annual mortality for a variety of diseases at four levels of cigarette smoking:
  1. never smoked
  2. 1-14 per day
  3. 15-24 per day
  4. > 25 per day

More information: ctsu.ox.ac.uk/research/british-doctors-study
## Simple linear regression: the data

<table>
<thead>
<tr>
<th>cigarettes smoked (per day)</th>
<th>CVD mortality (per 100,000 men per year)</th>
<th>lung cancer mortality (per 100,000 men per year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>572</td>
<td>14</td>
</tr>
<tr>
<td>10 (actually 1-14)</td>
<td>802</td>
<td>105</td>
</tr>
<tr>
<td>20 (actually 15-24)</td>
<td>892</td>
<td>208</td>
</tr>
<tr>
<td>30 (actually &gt;24)</td>
<td>1025</td>
<td>355</td>
</tr>
</tbody>
</table>

**CVD:** cardiovascular disease

Source: British Doctors’ Study
Simple linear regression: plots

Deaths for different smoking intensities

Quite tempting to conclude that cardiovascular disease deaths increase linearly with cigarette consumption...

```python
import pandas
import matplotlib.pyplot as plt

data = pandas.DataFrame({"cigarettes": [0,10,20,30],
    "CVD": [572,802,892,1025],
    "lung": [14,105,208,355]});
data.plot("cigarettes", "CVD", kind="scatter")
plt.title("Deaths for different smoking intensities")
plt.xlabel("Cigarettes smoked per day")
plt.ylabel("CVD deaths")
```
Aside: beware assumptions of causality

1964: the US Surgeon General issues a report claiming that cigarette smoking causes lung cancer, based mostly on correlation data similar to the previous slide.
1964: the US Surgeon General issues a report claiming that cigarette smoking causes lung cancer, based mostly on correlation data similar to the previous slide.

However, correlation is not sufficient to demonstrate causality. There might be some hidden genetic factor that causes both lung cancer and desire for nicotine.
Beware assumptions of causality

- To demonstrate the causality, you need a **randomized controlled experiment**

- Assume we have the power to force people to smoke or not smoke
  - and ignore moral issues for now!

- Take a large group of people and divide them into two groups
  - one group is obliged to smoke
  - other group not allowed to smoke (the “control” group)

- Observe whether smoker group develops more lung cancer than the control group

- We have eliminated any possible hidden factor causing both smoking and lung cancer

- More information: read about **design of experiments**
Fitting a linear model in Python

- In these examples, we use the `statsmodels` library for statistics in Python.
  - other possibility: the `scikit-learn` library for machine learning

- We use the formula interface to OLS regression, in `statsmodels.formula.api`

- Formulas are written as `outcome ~ observation`
  - meaning “build a linear model that predicts variable `outcome` as a function of input data on variable `observation`”
Fitting a linear model

```python
import numpy, pandas
import matplotlib.pyplot as plt
import statsmodels.formula.api as smf

df = pandas.DataFrame({
    "cigarettes": [0,10,20,30],
    "CVD": [572,802,892,1025],
    "lung": [14,105,208,355]
});
df.plot("cigarettes", "CVD", kind="scatter")

lm = smf.ols("CVD ~ cigarettes", data=df).fit()
xmin = df.cigarettes.min()
xmax = df.cigarettes.max()
X = numpy.linspace(xmin, xmax, 100)
# params[0] is the intercept (beta_0)
# params[1] is the slope (beta_1)
Y = lm.params[0] + lm.params[1] * X
plt.plot(X, Y, color="darkgreen")
```
Parameters of the linear model

- $\beta_0$ is the **intercept** of the regression line (where it meets the $X = 0$ axis)

- $\beta_1$ is the **slope** of the regression line

- Interpretation of $\beta_1 = 0.0475$: a “unit” increase in cigarette smoking is associated with a 0.0475 “unit” increase in deaths from lung cancer
Scatterplot of lung cancer deaths

Quite tempting to conclude that lung cancer deaths increase linearly with cigarette consumption...
Fitting a linear model

import numpy, pandas
import matplotlib.pyplot as plt
import statsmodels.formula.api as smf

df = pandas.DataFrame({"cigarettes": [0,10,20,30],
                      "CVD": [572,802,892,1025],
                      "lung": [14,105,208,355]});

df.plot("cigarettes", "lung", kind="scatter")

lm = smf.ols("lung ~ cigarettes", data=df).fit()

xmin = df.cigarettes.min()
xmax = df.cigarettes.max()

X = numpy.linspace(xmin, xmax, 100)

# params[0] is the intercept (beta_0)
# params[1] is the slope (beta_1)

Y = lm.params[0] + lm.params[1] * X

plt.plot(X, Y, color="darkgreen")

Download the associated Python notebook at risk-engineering.org
Using the model for prediction

**Q:** What is the expected lung cancer mortality risk for a group of people who smoke 15 cigarettes per day?

```python
import numpy, pandas
import statsmodels.formula.api as smf

df = pandas.DataFrame({"cigarettes": [0,10,20,30],
                      "CVD": [572,802,892,1025],
                      "lung": [14,105,208,355]})

# create and fit the linear model
lm = smf.ols(formula="lung ~ cigarettes", data=df).fit()

# use the fitted model for prediction
lm.predict({"cigarettes": [15]}) / 100000.0

# probability of mortality from lung cancer, per person per year
array([ 0.001705])
```
Assessing model quality

How do we assess how well the linear model fits our observations?
- make a visual check on a scatterplot
- use a quantitative measure of “goodness of fit”

Coefficient of determination $r^2$: a number that indicates how well data fit a statistical model
- it’s the proportion of total variation of outcomes explained by the model
- $r^2 = 1$: regression line fits perfectly
- $r^2 = 0$: regression line does not fit at all

For simple linear regression, $r^2$ is simply the square of the sample correlation coefficient $r$
Assessing model quality

I don't trust linear regressions when it's harder to guess the direction of the correlation from the scatter plot than to find new constellations on it.

Source: xkcd.com/1725, CC BY-NC licence
Information on the linear model

```python
> lm = smf.ols(formula='lung ~ cigarettes', data=df).fit()
> lm.summary()
```

**OLS Regression Results**

<table>
<thead>
<tr>
<th>Dep. Variable:</th>
<th>lung</th>
<th>R-squared:</th>
<th>0.987</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model:</td>
<td>OLS</td>
<td>Adj. R-squared:</td>
<td>0.980</td>
</tr>
<tr>
<td>Method:</td>
<td>Least Squares</td>
<td>F-statistic:</td>
<td>151.8</td>
</tr>
<tr>
<td>Date:</td>
<td>Wed, 06 Jan 2016</td>
<td>Prob (F-statistic):</td>
<td>0.00652</td>
</tr>
<tr>
<td>Time:</td>
<td>14:01:34</td>
<td>Log-Likelihood:</td>
<td>-16.359</td>
</tr>
<tr>
<td>No. Observations:</td>
<td>4</td>
<td>AIC:</td>
<td>36.72</td>
</tr>
<tr>
<td>Df Residuals:</td>
<td>2</td>
<td>BIC:</td>
<td>35.49</td>
</tr>
<tr>
<td>Df Model:</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Covariance Type:</td>
<td>nonrobust</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| coef | std err | t    | P>|t|  | [95.0% Conf. Int.] |
|------|---------|------|------|------------------|
| Intercept | 1.6000 | 17.097 | 0.094 | 0.934 | -71.964 75.164 |
| cigarettes | 11.2600 | 0.914 | 12.321 | 0.007 | 7.328 15.192 |

Omnibus: nan Durbin-Watson: 2.086
Prob(Omnibus): nan Jarque-Bera (JB): 0.534
Skew: -0.143 Prob(JB): 0.766
Kurtosis: 1.233 Cond. No. 31.4
Example: nosocomial infection risk

Longer stays in hospitals are associated with a higher risk of nosocomial infection.

Data source: SENIC survey on nosocomial risk
Example: blood pressure and BMI

Data on Body Mass Index and systolic blood pressure

A higher body mass index is correlated with higher blood pressure

Python with a Pandas dataframe:

```python
df.plot(x="BMI", y="SYBP", kind="scatter", alpha=0.5)
```

Example: blood pressure and BMI

Data on Body Mass Index and systolic blood pressure

A higher body mass index is correlated with higher blood pressure

Same data as previous slide, with a “hexplot” instead of scatterplot

Python with a Pandas dataframe:

```python
df.plot(x="BMI", y="SYBP", kind="hexbin", gridsize=25)
```

Example: blood pressure and BMI

Data on Body Mass Index and systolic blood pressure. A higher body mass index is correlated with higher blood pressure.

Same data as previous slide, with a kernel density plot instead of scatterplot.

Python with a Pandas dataframe using the Seaborn library:

```python
jointplot(data=df,
          x="BMI", y="SYBP",
          kind="kde")
```

Example: Intergenerational mobility in the USA

Percentage of children in college at age 19 plotted against the percentile rank of their parents’ income. Data for the USA.

Intergenerational mobility (for example chance of moving from bottom to top fifth of income distribution) is similar for children entering labor market today than in the 1970s. However, level of inequality has diminished, so consequences of the “birth lottery” are greater today.

(Political and moral implications of this analysis, and associated risks, are beyond the scope of these slides, but are one of our motivations for making these materials available for free...)

Data source: opportunityinsights.org/
Exercise: the “Dead grandmother problem”

**Problem.** Research by Prof. M. Adams suggests that the week prior to exam time is an extremely dangerous time for the relatives of university students. Data shows that a student’s grandmother is far more likely to die suddenly just before the student takes an exam, than at any other time of year.

**Theory.** Family members literally worry themselves to death over the outcome of their relatives’ performance on each exam.

**Task:** use linear regression to confirm that the severity of this phenomenon is correlated to the student’s current grade.

**Data source:** math.toronto.edu/mpugh/DeadGrandmother.pdf
Aside: linear regression in Excel

Summary:
- Functions SLOPE and INTERCEPT
- Correlation coefficient: function CORREL

Explanatory video: youtu.be/ExfknNCvBYg
In linear regression, the residual data is the difference between the observed data of the outcome variable $y$ and the predicted values $\hat{y}$

$$\text{residual} = y - \hat{y}$$

The residuals plot should look “random” (no discernible pattern)
- if the residuals are not random, they suggest that your model is systematically incorrect, meaning it can be improved
- see example to the right with no specific pattern

If you spot a trend in the residuals plot (increasing, decreasing, “U” shape), the data is most likely non-linear
- so a linear model is not a good choice for this problem...
Multivariate regression
What is multivariate linear regression?

Multivariate linear regression involves more than one predictor variable.
Recall the equation for univariate linear regression:

\[ \hat{y} = \beta_0 + \beta_1 x \]

Equation for **multivariate linear regression**:

\[ \hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \]

The outcome variable is assumed to be a linear combination of the predictor variables (the inputs)
Example: prediction using a multivariate dataset

- **Objective:** predict energy output at a Combined Cycle Power Plant

- **Data available:** hourly averages of variables

<table>
<thead>
<tr>
<th>Meaning</th>
<th>Name</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambient Temperature</td>
<td>AT</td>
<td>1.81 – 37.11°C</td>
</tr>
<tr>
<td>Ambient Pressure</td>
<td>AP</td>
<td>992.89 – 1033.30 millibar</td>
</tr>
<tr>
<td>Relative Humidity</td>
<td>RH</td>
<td>25.56% – 100.16%</td>
</tr>
<tr>
<td>Exhaust Vacuum</td>
<td>V</td>
<td>25.36 – 81.56 cm Hg</td>
</tr>
<tr>
<td>Net hourly electrical energy output</td>
<td>PE</td>
<td>420.26 – 495.76 MW</td>
</tr>
</tbody>
</table>

- Let’s try to build a multivariate linear model to predict PE given inputs AT, AP, RH and V

Dataset: archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant
Example: prediction using a multivariate dataset

- Dataset contains 9568 data points collected from a combined cycle power plant over 6 years, when power plant was under full load.

- A combined cycle power plant is composed of gas turbines, steam turbines and heat recovery steam generators.
  - Electricity is generated by gas & steam turbines, which are combined in one cycle.
  - Three ambient variables affect performance of the gas turbine.
  - Exhaust vacuum affects performance of the steam turbine.

- Data consists of hourly averages taken from various sensors located around the plant that record the ambient variables every second.

- Let's load it into Python and examine it using the `pandas` library.

Dataset: archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant
Example: prediction using a multivariate dataset

```python
import pandas
data = pandas.read_csv("data/CCPP.csv")
data.head()

<table>
<thead>
<tr>
<th>AT</th>
<th>V</th>
<th>AP</th>
<th>RH</th>
<th>PE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>14.96</td>
<td>41.76</td>
<td>1024.07</td>
<td>73.17</td>
</tr>
<tr>
<td>1</td>
<td>25.18</td>
<td>62.96</td>
<td>1020.04</td>
<td>59.08</td>
</tr>
<tr>
<td>2</td>
<td>5.11</td>
<td>39.40</td>
<td>1012.16</td>
<td>92.14</td>
</tr>
<tr>
<td>3</td>
<td>20.86</td>
<td>57.32</td>
<td>1010.24</td>
<td>76.64</td>
</tr>
<tr>
<td>4</td>
<td>10.82</td>
<td>37.50</td>
<td>1009.23</td>
<td>96.62</td>
</tr>
</tbody>
</table>

data.describe()

<table>
<thead>
<tr>
<th>AT</th>
<th>V</th>
<th>AP</th>
<th>RH</th>
<th>PE</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>9568.000000</td>
<td>9568.000000</td>
<td>9568.000000</td>
<td>9568.000000</td>
</tr>
<tr>
<td>mean</td>
<td>19.651231</td>
<td>54.305804</td>
<td>1013.259078</td>
<td>73.308978</td>
</tr>
<tr>
<td>std</td>
<td>7.452473</td>
<td>12.707893</td>
<td>5.938784</td>
<td>14.600269</td>
</tr>
<tr>
<td>min</td>
<td>1.810000</td>
<td>25.360000</td>
<td>992.890000</td>
<td>25.560000</td>
</tr>
<tr>
<td>25%</td>
<td>13.510000</td>
<td>41.740000</td>
<td>1009.100000</td>
<td>63.327500</td>
</tr>
<tr>
<td>50%</td>
<td>20.345000</td>
<td>52.080000</td>
<td>1012.940000</td>
<td>74.975000</td>
</tr>
<tr>
<td>75%</td>
<td>25.720000</td>
<td>66.540000</td>
<td>1017.260000</td>
<td>84.830000</td>
</tr>
<tr>
<td>max</td>
<td>37.110000</td>
<td>81.560000</td>
<td>1033.300000</td>
<td>100.160000</td>
</tr>
</tbody>
</table>
```

Dataset: archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant
We can obtain a first impression of the dependency between variables by examining a multidimensional scatterplot.

```python
from pandas.tools.plotting import scatter_matrix
data = pandas.read_csv("data/CCPP.csv")
scatter_matrix(data, diagonal="kde")
```

In this matrix, the diagonal contains a plot of the distribution of each variable.
Interpreting the scatterplot matrix

Observations:
▷ approximately linear relationship between PE and the negative of AT
▷ approximately linear relationship between PE and negative of V
Visualizing multivariate data: 3D plotting

It is sometimes useful to examine 3D plots of your observations

```python
from mpl_toolkits.mplot3d import Axes3D
import matplotlib.pyplot as plt

fig = plt.figure(figsize=(12, 8))
ax = Axes3D(fig, azim=-115, elev=15)
ax.scatter(data['AT'], data['V'], data['PE'])
ax.set_xlabel("AT")
ax.set_ylabel("V")
ax.set_zlabel("PE")
```
Importance of preliminary data analysis

Consider a study that measures weekly exercise and cholesterol in various age groups.

If we plot exercise against cholesterol and segregate by age, we see a downward trend in each group: more exercise leads to lower cholesterol.
Importance of preliminary data analysis

If we don’t segregate by age, we get the plot to the right, which could lead to an incorrect conclusion that more exercise is correlated with more cholesterol.

There is an underlying variable age: older people tend to exercise more, and also have higher cholesterol.

This example from *Causal inference in statistics: a primer*, J. Pearl et al, 2016
**CCPP example: least squares regression with Python**

```python
# create fitted model using “formula” API of the statsmodels library
import statsmodels.formula.api as smf
> lm = smf.ols(formula='PE ~ AT + V + AP + RH', data=data).fit()
> print(lm.params)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>451.067793</td>
</tr>
<tr>
<td>AT</td>
<td>-1.974731</td>
</tr>
<tr>
<td>V</td>
<td>-0.234992</td>
</tr>
<tr>
<td>AP</td>
<td>0.065540</td>
</tr>
<tr>
<td>RH</td>
<td>-0.157598</td>
</tr>
</tbody>
</table>
```

This means that the best formula to estimate output power as a function of AT, V, AP and RH is

\[
PE = 451.067793 - 1.974731 \times AT - 0.234992 \times V + 0.065540 \times AP - 0.157598 \times RH
\]
The residuals for each predictor variable look random, except for a mild quadratic shape for AT, which we will ignore here.
One assumption underlying linear regression is that the variance of the residuals is normally distributed (follows a Gaussian distribution).

Can be checked by plotting a histogram or a Q-Q plot of the residuals, as shown to the right.

Example to the right: we have a deviation from normality for large prediction errors, but overall residuals follow a normal distribution.

Download the associated Python notebook at risk-engineering.org
CCPP example: prediction

Assuming the values below for our input variables, what is the predicted output power?

<table>
<thead>
<tr>
<th>AT</th>
<th>9.48</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>44.71</td>
</tr>
<tr>
<td>AP</td>
<td>1019.12</td>
</tr>
<tr>
<td>RH</td>
<td>66.43</td>
</tr>
</tbody>
</table>

> m = pandas.DataFrame({"AT": [9.48], "V": [44.71], "AP": [1019.12], "RH": [66.43]})
> lm.predict(m)
[ 478.25471442]

Conclusion: the predicted output power is 478.3 MW.
Assessing goodness of fit: $R^2$

For multiple linear regression, the coefficient of determination $R^2$ is calculated as

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where
- $SS_{res} = \sum_i (y_i - \tilde{y}_i)^2$ is the sum of the square of the residuals
- $SS_{tot} = \sum_i (y_i - \bar{y})^2$ is the total sum of squares
- $y_i$ are the observations, for $i = 1...n$
- $\tilde{y}_i$ are the predictions, for $i = 1...n$
- $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ is the mean of the observations

- The better the fit, the closer $R^2$ is to 1

- $R^2$ measures the proportion of variance in the observed data that is explained by the model

Areas of red squares: squared residuals with respect to the average value

Areas of blue squares: squared residuals with respect to the linear regression
Determining $R^2$ in Python

```python
> lm.summary()

OLS Regression Results
==============================================================================
Dep. Variable: PE R-squared: 0.927
Model: OLS Adj. R-squared: 0.927
Method: Least Squares F-statistic: 2.295e+04
Date: Tue, 05 Jan 2016 Prob (F-statistic): 0.00
No. Observations: 7196 AIC: 4.234e+04
Df Residuals: 7191 BIC: 4.238e+04
Df Model: 4
Covariance Type: nonrobust
==============================================================================

| coef | std err | t     | P>|t|  | [95.0% Conf. Int.] |
|------|---------|-------|------|------------------|
| Intercept | 460.9650 | 11.308 | 40.764 | 0.000 | 438.798 483.132 |
| AT    | -1.9809  | 0.018  | -111.660 | 0.000 | -2.016 -1.946 |
| V     | -0.2303  | 0.008  | -27.313 | 0.000 | -0.247 -0.214 |
| AP    | 0.0556   | 0.011  | 5.073  | 0.000 | 0.034 0.077 |
| RH    | -0.1576  | 0.005  | -32.827 | 0.000 | -0.167 -0.148 |

Omnibus: 864.810 Durbin-Watson: 2.009
Prob(Omnibus): 0.000 Jarque-Bera (JB): 4576.233
Skew: -0.459 Prob(JB): 0.00
Kurtosis: 6.797 Cond. No. 2.13e+05
```
Warnings concerning linear regression
Warnings concerning use of linear regression

1. Check that your data is really linear!

2. Make sure your sample size is sufficient

3. Don’t use a regression model to predict responses outside the range of data that was used to build the model

4. Results can be highly sensitive to treatment of outliers

5. Multiple regression: check that your predictors are independent

6. Beware order of effect problems
   • regression shows correlation but does not necessarily imply causality

7. Beware the regression to the mean effect
Check assumptions underlying linear regression

- Examine scatterplot of outcome variable with each predictor to validate the assumption of linearity

- Other assumptions underlying the use of linear regression:
  - Check that the mean of the residuals is almost equal to zero for each value of outcome
  - Check that the residuals have constant variance (→ residuals scatterplot on slide 24)
  - Check that residuals are uncorrelated (→ residuals scatterplot)
  - Check that residuals are normally distributed (→ residuals histogram or QQ-plot) or that you have an adequate sample size to rely on large sample theory
Make sure your sample size is sufficient

- There are no rules on required sample size for a regression analysis
  - depends on the number of predictor variables, on the effect size, the objective of the analysis

- Some general observations:
  - bigger samples are better (give more confidence in the model)
  - sample size is often determined by pragmatic considerations (measurements may be expensive, limited historical data available)
  - sample size should be seen as one consideration in an optimization problem where the cost in time/money/effort of obtaining more data is weighed against the benefits (better predictions, improved understanding)
Extrapolate with care

When using a linear model for prediction, be very careful when predicting responses outside of the range of data that was used to build the model.

Make sure you have well-grounded scientific reasons for arguing that the model also applies in areas where you don’t have available data.

Image source: xkcd.com/605/, CC BY-NC licence

To infinity and beyond!
Treatment of outlier data

▷ Real datasets often contain **spurious data points**
  • errors made in measurement, noise, data entry errors...

▷ These may have a significant impact on your predictions

▷ However, some outlier data may just be “different” but meaningful observations
  • possibly an early warning sign of an upcoming catastrophe!

▷ The best method of handling outliers depends on the objective of your analysis, on how you obtained your data...
Illustration: in the early 1980s, scientists were shocked by a dramatic seasonal drop in ozone levels over Antarctica and spent two years analyzing their data. Satellites had been correctly recording the “hole in the ozone layer”, but were programmed to reject outliers as anomalies.

[R. Benedick, Scientific American, April 1992]
Recommendations for handling outliers

▷ Analyze outliers individually, for instance by plotting your data

▷ Eliminate from the dataset any outliers that you are confident you can identify as being the result of errors in measurement or data entry

▷ For remaining outliers, report prediction results both with and without the outliers

▷ Consider using a robust linear regression technique
  • example: RLM from the statsmodels library (Python)
  • RANSAC from the scikit-learn library (Python)
Beware of order of effect problems

Consider infant mortality data from the World Bank

Data source: data.worldbank.org
Two variables:
- infant mortality rate (per 1000 births)
- number of births per woman

Which is the predictor variable and which is the outcome?
Two variables:
- infant mortality rate (per 1000 births)
- number of births per woman

Which is the predictor variable and which is the outcome?

Choice 1: fertility = f(infant-mortality)
- predictor: infant mortality rate
- outcome: births per woman

Choice 2: infant-mortality = f(fertility)
- predictor: births per woman
- outcome: infant mortality rate

Data source: data.worldbank.org
The answer depends on the **framing of the research question**

If hypothesis is influence of infant mortality on number of births per woman, then
- predictor: infant mortality rate
- outcome: births per woman

If hypothesis is influence of number of births per woman on infant mortality, then
- predictor: births per woman
- outcome: infant mortality rate
Directionality of effect problem

Examples:

- People who exercise more tend to have better health
- Police departments with higher budgets tend to be located in areas with high crime levels
- Middle-aged men who wear hats are more likely to be bald
- Young smokers who buy contraband cigarettes tend to smoke more
Regression to the mean

- Following an extreme random event, the next random event is likely to be less extreme
  - if a variable is extreme on its first measurement, it will tend to be closer to the average on its second measurement

- Examples:
  - If today is extremely hot, you should probably expect tomorrow to be hot, but not quite as hot as today
  - If a baseball player just had by far the best season of his career, his next year is likely to be a disappointment

- Extreme events tend to be followed by something closer to the norm
Statistical regression to the mean predicts that patients selected for abnormalcy will, on the average, tend to improve. We argue that most improvements attributed to the placebo effect are actually instances of statistical regression.

Thus, we urge caution in interpreting patient improvements as causal effects of our actions and should avoid the conceit of assuming that our personal presence has strong healing powers. [McDonald et al 1983]

Group of patients that are treated with a placebo are affected by two processes:

- genuine psychosomatic placebo effect
- “get better anyway” effect (regression to the mean)

Regression to the mean

▷ Classical example of regression to the mean: **effectiveness of speed cameras** in preventing accidents

▷ Speed cameras tend to be installed after an exceptional series of accidents at that location

▷ If the accident rate is particularly high somewhere one year, it will probably be lower the next year
  • irrespective of whether a speed camera is installed...

▷ To avoid this bias, implement a *randomized trial*
  • choose several similar sites
  • allocate them at random to have a camera or no camera
  • check whether the speed camera has a statistically measurable effect
Aside: origin of the term

▷ Francis Galton (1822–1911) was an English anthropologist and statistician (and polymath)

▷ Discovered/formalized the statistical concept of correlation

▷ Collected data on the height of the descendants of extremely tall and extremely short trees
  • to analyze how “co-related” trees were to their parents
  • publication: *Regression Towards Mediocrity in Hereditary Stature* (1866)

“It appeared from these experiments that the offspring did not tend to resemble their parents seeds in size, but to be always more mediocre than they - to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were small.

▷ But towards the end of his life, studied whether human ability was hereditary and promoted eugenics...
Linear regression techniques are not applicable for category data, such as success/failure data
• use generalized linear models (GLM) instead

Sometimes **machine learning** algorithms can be more appropriate than regression techniques
• example algorithms: random forest, support vector machines, neural networks
Further reading

▷ The Stanford Online class on Statistical Learning introduces supervised learning with a focus on regression and classification methods
   → online.stanford.edu

▷ The online, open-access textbook *Forecasting: principles and practice*
   → otexts.org/fpp2 (uses R rather than Python)

▷ Online book *Practical regression and Anova using R*
   → cran.r-project.org/doc/contrib/Faraway-PRA.pdf

For more free content on risk engineering, visit risk-engineering.org
Feedback welcome!

Was some of the content unclear? Which parts were most useful to you? Your comments to feedback@risk-engineering.org (email) or @LearnRiskEng (Twitter) will help us to improve these materials. Thanks!

For more free content on risk engineering, visit risk-engineering.org