

Assignment Project Exam Help

Predictive Analytics

Week 2: Linear Regression and Statistical Thinking

<https://powcoder.com>

Semester 2, 2018

Discipline of Business Analytics, The University of Sydney Business School

Add WeChat powcoder

Assignment Project Exam Help

1. Statistical and Machine Learning foundations and applications.

2. Advanced regression methods.

3. Classification methods.

4. Time series forecasting.

<https://powcoder.com>

Add WeChat powcoder

Before the lecture 2, review linear algebra, especially matrix multiplication, rank, determinant and inverse.

Week 2: Linear Regression and Statistical Thinking

1. Introduction

Assignment Project Exam Help

2. The least squares algorithm

3. The MLR model

<https://powcoder.com>

4. Statistical properties

Add WeChat powcoder

5. Interpreting a linear regression model

6. Regression modelling

Assignment Project Exam Help

Introduction
<https://powcoder.com>

Add WeChat powcoder

Linear regression

The linear regression is a simple and widely used method for supervised learning. There are several important reasons for developing an in-depth understanding of this method.

Assignment Project Exam Help

- It is very useful for prediction in many settings.
- It is extremely useful conceptually. Many advanced statistical learning methods can be understood as extensions and generalisations of linear regression.
- Due to its simplicity, linear regression is often a useful jumping-off point for model building and analysis.
- Interpretability.

<https://powcoder.com>

Add WeChat powcoder

Example: advertisement data

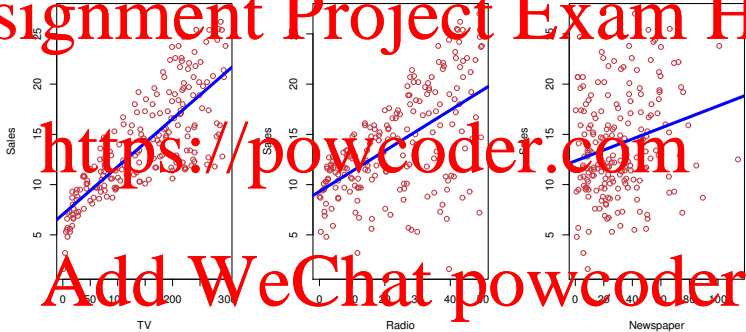
Consider from example the advertisement data from the ISL

textbook (see next slide). Possible questions

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we predict future sales?
- Is there synergy among the advertisement media?

Example: advertisement data

Assignment Project Exam Help



(Figure from ISL)

Example: advertisement data

Assignment Project Exam Help

To answer our questions, we can use a model such as

<https://powcoder.com>

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper} + \varepsilon$$

Add WeChat powcoder

Statistical thinking

Statistical thinking is using statistical models, statistical theory, and critical thinking to learn from data.

Assignment Project Exam Help

- How do I design a study to answer a certain question?

- How relevant and representative are my data?

- What is the variability in my data? Can I reliably draw conclusions in light of this variability?

- How do I correctly interpret my results?

- Can I generalise my conclusions in the way that I would like to?

- What are the limitations of my analysis?

<https://powcoder.com>

Add WeChat powcoder

Assignment Project Exam Help

~~The least squares algorithm~~
<https://powcoder.com>

Add WeChat powcoder

Assignment Project Exam Help

In the linear regression method for prediction, we consider a regression function of the form

$$f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

We learn the prediction coefficients $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ by fitting the model to the training data using the least squares method.

Least squares

Let $\mathcal{D} = \{(y_i, \mathbf{x}_i)\}_{i=1}^N$ be the training data. We define the **residual sum of squares** as a function of parameter values β as

$$\text{RSS}(\beta) = \sum_{i=1}^N (y_i - f(\mathbf{x}_i; \beta))^2$$

$$= \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2$$

Assignment Project Exam Help

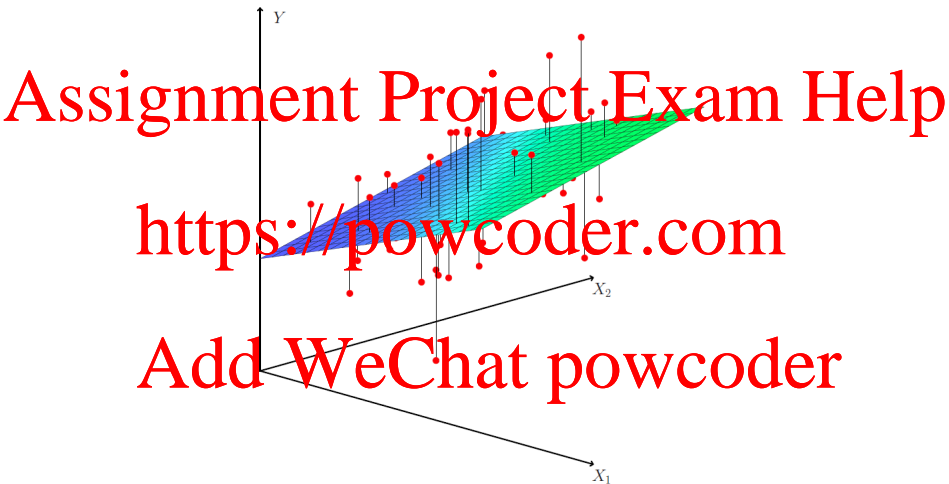
The ordinary least squares (OLS) method selects the coefficient values that minimise the residual sum of squares

<https://powcoder.com>

$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2$

Add WeChat powcoder

Least squares



(Figure from ISL)

Assignment Project Exam Help

If our loss function $L(y, f(x))$ is the squared error loss, the OLS algorithm consists of minimising the empirical loss for our choice of predictive function:

<https://powcoder.com>

$\hat{\beta} = \operatorname{argmin}_{\beta} \frac{1}{N} \sum_{i=1}^N L(y_i, f(\mathbf{x}_i))$
Add WeChat powcoder

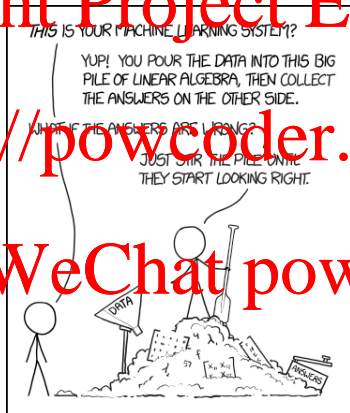
Least squares and linear algebra

In order to obtain a solution to the OLS minimisation problem, we need linear algebra.

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder



Assignment Project Exam Help

<https://powcoder.com>

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \dots & x_{Np} \end{pmatrix}$$

Add WeChat powcoder

We equivalently write the RSS as

Assignment Project Exam Help

<https://powcoder.com>

$$\begin{aligned}\text{RSS}(\beta) &= (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \\ &= \mathbf{y}^T \mathbf{y} - 2\beta^T \mathbf{X}^T \mathbf{y} + \beta^T \mathbf{X}^T \mathbf{X} \beta\end{aligned}$$

Add WeChat powcoder

We optimise the RSS by taking the $p+1$ partial derivatives and setting them to zero.

Vector differentiation rules

Assignment Project Exam Help

Let x and a be vectors of equal dimension and A a matrix with column dimension the same as number of rows in x . Then:

$$\frac{d(x^T a)}{dx} = a$$

Add WeChat powcoder

$$\frac{d(x^T A x)}{dx} = (A + A^T)x$$

Assignment Project Exam Help

$$RSS(\beta) = \mathbf{y}^T \mathbf{y} - 2\mathbf{y}^T \mathbf{X} \beta + \beta^T \mathbf{X}^T \mathbf{X} \beta$$

The vector of partial derivatives is

$$\begin{aligned} \frac{d(RSS(\beta))}{d\beta} &= \frac{d(\mathbf{y}^T \mathbf{y})}{d\beta} - \frac{d(2\mathbf{y}^T \mathbf{X} \beta)}{d\beta} + \frac{d(\beta^T \mathbf{X}^T \mathbf{X} \beta)}{d\beta} \\ &= \mathbf{0} - 2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \beta \end{aligned}$$

OLS estimates

The first order condition is:

$$\frac{d(\text{RSS}(\beta))}{d\beta} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\beta = 0$$

The least squares estimate $\hat{\beta}$ therefore satisfies

$$\mathbf{X}^T\mathbf{X}\hat{\beta} = \mathbf{X}^T\mathbf{y}$$

If $(\mathbf{X}^T\mathbf{X})^{-1}$ is invertible, left multiplication with $(\mathbf{X}^T\mathbf{X})^{-1}$ gives the unique solution

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

OLS for big data?

Assignment Project Exam Help

The OLS solution is $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$, given that you can compute the matrix $\mathbf{X}^T \mathbf{X}$.

\mathbf{X} is a matrix of size $N \times (p+1)$. If N is very large, then \mathbf{X} is so large that it is impossible to compute $\mathbf{X}^T \mathbf{X}$ or it is close to being singular.

For big data, it's challenging to compute this matrix! Solutions?

<https://powcoder.com>
Add WeChat powcoder

$X^T X$ non-invertible?

- Reason: Multicollinearity problem or redundant predictors
Rank and determinant of $X^T X = ?$
 - Solution: Drop one or more highly correlated predictors from the model or collect more data.
- Reason: The number of predictors is too large (e.g., $N \ll p$).
Rank and determinant of $X^T X = ?$
 - Solution: Drop some predictors or collect more data; Add regularization term into the model. More details later.
- For real matrices X :
$$\text{rank}(X^T X) = \text{rank}(X X^T) = \text{rank}(X) = \text{rank}(X^T)$$

Fitted values

The fitted values based on the training inputs are

Assignment Project Exam Help

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_{ij}$$

The vector of fitted values for the entire sample is:

$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

We refer to $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ as the **hat matrix**.

Residuals

The regression **residuals** are:

Assignment Project Exam Help

$$e_i = y_i - \hat{y}_i \\ = y_i - \hat{\beta}_0 - \sum_{j=1}^p \hat{\beta}_j x_{ij}$$

<https://powcoder.com>

The vector of residuals is:

Add WeChat powcoder

$$\begin{aligned} \mathbf{e} &= \mathbf{y} - \hat{\mathbf{y}} \\ &= \mathbf{y} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \\ &= \left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right) \mathbf{y}. \end{aligned}$$

Measuring fit

We can show that

Assignment Project Exam Help

$$TSS = RegSS + RSS$$

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

<https://powcoder.com>

- TSS: total sum of squares.
- RegSS: regression sum of squares.
- RSS: residual sum of squares.

Measuring fit

$$R^2 = \frac{\text{RegSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Assignment Project Exam Help

Interpretation:

- The R^2 measures the proportion of the variation in the response data that is accounted for by the estimated linear regression model.
- The R^2 can only increase when you add another variable to the model.
- The R^2 is an useful part of the regression toolbox, but it does not measure the predictive accuracy of the estimated regression, or more generally how good the model is.

Assignment Project Exam Help

Let $\hat{\beta}$ be the OLS coefficients obtained from the training sample.

<https://powcoder.com>

$$\hat{y}_0 = \hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_{0j}$$

Add WeChat powcoder

Assignment Project Exam Help

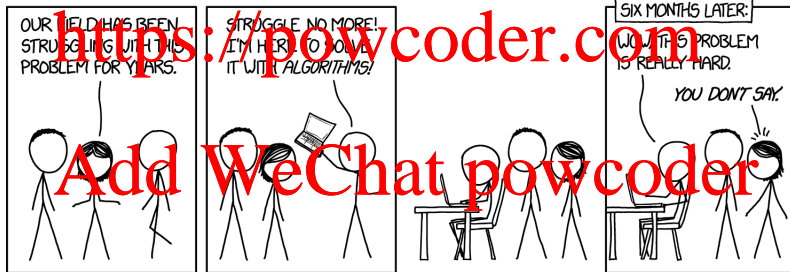
The MLR model
<https://powcoder.com>

Add WeChat powcoder

Models and algorithms

So far, we have talked about the least squares algorithm and even arrived at predictions without reference to a model. The current

practice of data science places large emphasis on algorithmic thinking towards problem solving.



<https://xkcd.com/1831/>

Statistical models

A **statistical model** is a description of a data generating process based on a set of mathematical assumptions about the population and the sampling process.

A **regression model** is a description of the relationship between a response variable Y and predictors X_1, \dots, X_p . More formally, it is a model of the form $p(y|x; \theta)$.

Add WeChat powcoder

Formulating statistical models and making assumptions allow us to say more about a problem.

The Multiple Linear Regression (MLR) model

1. Linearity: if $X = \mathbf{x}$, then

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon$$

for some population parameters $\beta_0, \beta_1, \dots, \beta_p$ and a random error ε .

2. The conditional mean of ε given X is zero: $E(\varepsilon|X) = 0$.
3. Constant error variance: $\text{Var}(\varepsilon|X) = \sigma^2$.

4. Independence: all the error pairs ε_i and ε_j ($i \neq j$) are independent.

5. The distribution of X_1, \dots, X_p is arbitrary.

6. There is no perfect multicollinearity.

Checking the assumptions

It is fundamental to check the assumptions with data. We do this with **residual diagnostics**. The following plots are often useful:

Assignment Project Exam Help

- Fitted values against residuals.

- Predictors against residuals.

<https://powcoder.com>

- Fitted values against squared or absolute residuals.

- Predictors against squared or absolute residuals.

Add WeChat powcoder

- Residual distribution.

- If the observations are ordered: residuals against coordinates (time and/or space).

Assignment Project Exam Help

Statistical properties
<https://powcoder.com>

Add WeChat powcoder

Sampling distribution of an estimator

Assignment Project Exam Help

In classical statistics, the population parameter β is fixed and the data is a random sample from the population. We estimate β by applying an **estimator** $\hat{\beta}(\mathcal{D})$ to data (in our case the OLS algorithm).

We study the uncertainty of an estimate by computing the **sampling distribution** of the estimator.

Sampling distribution of an estimator

Assignment Project Exam Help

Imagine that we draw many different datasets $\mathcal{D}^{(s)}$ ($s = 1, \dots, S$) from the true model $p(\mathbf{y}|\mathbf{X}; \beta)$. Each dataset has size N .

For each of these datasets, we apply the estimator $\hat{\beta}(\cdot)$ and obtain a set of estimates $\{\hat{\beta}(\mathcal{D}^{(s)})\}$. The sampling distribution is the induced distribution on $\hat{\beta}(\cdot)$ as $S \rightarrow \infty$.

Add WeChat powcoder

This concept is not necessarily intuitive since it refers to hypothetical datasets rather than data that we do have.

Sampling distribution of an estimator

Establishing the sampling distribution allows us to answer questions such as:

- Is there a significant relationship between the response and the predictors?
- Are all the predictors related to the response, or only a subset?
- How accurate are our coefficient estimates?
- How accurate are our predictions?

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Sampling distribution

Under the Gaussian MLR model with $\varepsilon \sim N(0, \sigma^2)$, we can obtain an exact sampling distribution for the OLS estimator,

Assignment Project Exam Help

$$\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}^T \mathbf{X})^{-1})$$

When estimating σ^2 , we have

<https://powcoder.com>

$$\frac{\hat{\beta}_j - \beta_j}{\text{SE}(\hat{\beta}_j)} \sim ?$$

We can then rely on this distribution for hypothesis testing.

Add WeChat powcoder

Review your study notes of previous units or the reference book for: OLS estimator sample distribution, regression coefficient significance testing, confidence interval, ANOVA, etc.

Assignment Project Exam Help

Interpreting a linear regression

model
<https://powcoder.com>

Add WeChat powcoder

Advertisement data

We now estimate the linear regression model

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper} + \varepsilon$$

Assignment Project Exam Help

To interpret the results, we need to note the following:

- <https://powcoder.com>
The observational units in the data are markets.

- The response variable (sales) is in thousands of units.

- The predictors are in thousands of dollars.

Add WeChat powcoder

What is the population of interest? (You always need to be able to answer this question)

Advertisement data

OLS Regression Results

```
=====
Dep. Variable:          Sales    R-squared:                0.897
Model:                  OLS      Adj. R-squared:            0.896
Method:                 Least Squares    F-statistic:          170.3
Date:                   Prob (F-statistic):    1.50e-96
Time:                   Log-Likelihood:        -386.18
No. Observations:      200      AIC:                  780.4
Df Residuals:          196      BIC:                  793.6
Df Model:               3
```

```
Covariance Type:        nonrobust
=====
              coef      std err          t      P>|t|      [0.025      0.975]
```

```
-----
Intercept      2.9389      0.312      9.422      0.000      2.324      3.554
TV              0.0458      0.001     32.809      0.000      0.043      0.049
Radio          0.0875      0.009     21.892      0.000      0.072      0.096
Newspaper     -0.0010      0.006     -0.177      0.860     -0.013      0.011
=====
```

```
Omnibus:          60.414    Durbin-Watson:          2.084
Prob(Omnibus):    0.000    Jarque-Bera (JB):       151.241
Skew:             -1.327    Prob(JB):               1.44e-33
Kurtosis:         6.332    Cond. No.               454.
=====
```

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Interpreting coefficients

$$\text{sales} = \underset{(0.312)}{-2.9389} + \underset{(0.001)}{0.0458 \times \text{TV}} + \underset{(0.009)}{0.1885 \times \text{radio}} - \underset{(0.006)}{0.0010 \times \text{newspaper}}$$

<https://powcoder.com>
Interpretation (TV):

Add WeChat powcoder
If we select two markets from the population, where the radio and newspaper budgets are the same, but the TV budget differs by 100 dollars, we would expect 4.58 more units sold in the market with higher TV budget.

Interpreting coefficients

Mathematically:

$$\beta_j = E(Y|X_1 = x_1, \dots, X_{j-1} = x_{j-1}, X_{j+1} = x_{j+1}, \dots, X_p = x_p) - E(Y|X_1 = x_1, \dots, X_{j-1} = x_{j-1}, X_{j+1} = x_{j+1}, \dots, X_p = x_p)$$

For example, with $p = 2$ and focusing on the first predictor:

$$\begin{aligned} & E(Y|X_1 = x_1 + 1, X_2 = x_2) - E(Y|X_1 = x_1, X_2 = x_2) \\ &= E[\beta_0 + \beta_1(x_1 + 1) + \beta_2x_2 + \varepsilon] - E[\beta_0 + \beta_1x_1 + \beta_2x_2 + \varepsilon] \\ &= [\beta_0 + \beta_1(x_1 + 1) + \beta_2x_2] - [\beta_0 + \beta_1x_1 + \beta_2x_2] \\ &= \beta_1 \end{aligned}$$

Omitted variables

With observational data, the assumption that $E(\varepsilon|X = \mathbf{x}) = 0$ is generally not satisfied. In this case, there are **omitted variables**, variables that are correlated with both the predictor and the response. This leads to **omitted variable bias** when estimating regression coefficients.

Here is an example: if we regress wealth on the number of luxury cars owned, the slope is positive (luxury cars predict wealth). However, we can imagine that buying more luxury cars will not make you richer.

Example: education and wages

OLS Regression Results

```
=====
Dep. Variable:          Hourly wage    R-squared:                0.162
Model:                  OLS            Adj. R-squared:          0.161
Method:                 Least Squares   F-statistic:              729
Date:                   Prob (F-statistic): 0.00
Time:                   Log-Likelihood: -57425.
No. Observations:      17919          AIC:                    1.149e+05
Df Residuals:          17916          BIC:                    1.149e+05
Df Model:               2
Covariance Type:       no robust
=====
```

```
=====
              coef      std err          t      P>|t|      [0.025      0.975]
-----
Intercept    -7.5017      0.337     -22.278      0.000     -8.162     -6.842
Education     1.1977      0.024     50.255      0.000      1.147      1.240
Experience     0.5111      0.011     46.372      0.000      0.429      0.493
=====
```

```
Omnibus:            10774.032    Durbin-Watson:           0.744
Prob(Omnibus):      0.000        Jarque-Bera (JB):        237446.384
Skew:               2.484        Prob(JB):                0.00
Kurtosis:           20.128        Cond. No.               117.
=====
```

Causal analysis

Causal analysis means to estimate a model of the type $E(Y|\text{do } X = x)$. This is an explicit intervention. “If we do $X = x$, then we predict $E(Y|X = x)$ ”.

<https://powcoder.com>

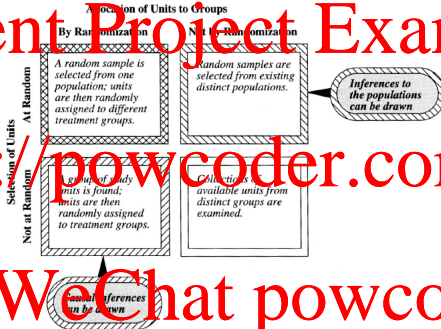
This is different from predictive modelling: “if we observe $X = x$, then we predict $E(Y|X = x)$ ”.

Add WeChat powcoder

Causal analysis requires an appropriate **study design** (such as A/B testing).

Study designs

Display 1.5 Statistical inferences permitted by study designs



Ramsey and Shafter (2002).

Assignment Project Exam Help

- For our purposes, the textbook is not sufficiently rigorous regarding the interpretation of linear regression coefficients.

<https://powcoder.com>

- While our interpretation is less simple than the one provided by most textbooks, it is the correct one for observational data that is prevalent in business.

Add WeChat powcoder

Assignment Project Exam Help

Regression modelling
<https://powcoder.com>

Add WeChat powcoder

Regression modelling

- All the material from Statistical Modelling for Business continues to be relevant in Predictive Analytics.

Assignment Project Exam Help

- In particular, constructing a helpful set of predictor variables is extremely important for supervised learning as it is often essential to improving performance. Constructing a helpful set of predictor variables possible is extremely important for supervised learning. This is known in machine learning and data science as **feature engineering**.. This is known in machine learning and data science as **feature engineering**.

- It is also useful to build models that fit as much as possible the assumptions on data (for example, constant error variance).

- Data transformation (in particular log and power transformations).

Assignment Project Exam Help

- Categorical predictors.

- <https://powcoder.com>

- Polynomial regression.

- Add WeChat powcoder

- Regression splines.

- Robust regression.

Potential problems

- Nonlinearity.

- Non-constant error variance.

- Correlated errors.

- Outliers and high leverage points.

- Multicollinearity.

- Non-Gaussianity.

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Review questions

- How do we obtain the OLS estimates? Go through the full process.

Assignment Project Exam Help

- What is a sampling distribution?

- We formulated several questions about the advertisement data. Answer some of these questions based on the Python output in the slides.

<https://powcoder.com>

- What is the correct interpretation of a linear regression model coefficient with observational data?
- What is the difference between predictive and causal analysis?

Add WeChat powcoder