目录

[Polling 1](#_Toc174637571)

[Correlations 2](#_Toc174637572)

[Regression 4](#_Toc174637573)

[Multiple Regression 7](#_Toc174637574)

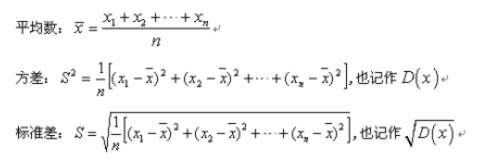
[Evaluating Regression Models 8](#_Toc174637575)

[Transforming Variables in Regression 16](#_Toc174637576)

[Spatial Models 19](#_Toc174637577)

[Spatial Regression Analysis 21](#_Toc174637578)

# Polling



Standard Error 标准误差：随机变量的概率密度函数的标准差，特别是指从具有正态分布的总体中抽取的样本的平均值的标准误差，它等于正态分布的标准差除以样本大小的平方根。所以样本数量越大，标准误差越小。

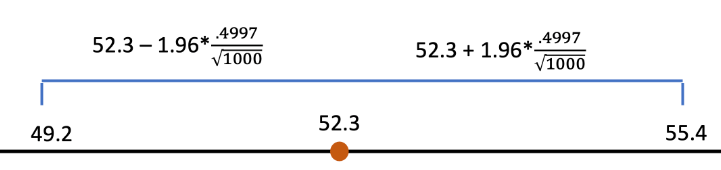
Confidence Intervals 置信区间：统计学中用于估计一个参数范围的区间，表示对于参数估计值的置信度。

Central Limit Theorem 中心极限定理：概率论和统计学中的几个基本定理之一，它阐述了在一定条件下，独立随机变量之和的分布可以近似为正态分布。

we can expect that result to be within 1 standard error of the population mean 68% of the time. We can expect it to be within 2 standard error of the population mean 95% of the time. And we can expect it to be within 3 standard errors 99.7% of the time.



因此当Z为2（准确来讲是1.96）时，总体均值有95%的概率位于该置信区间内。



R获得均值mean(election$AnticipateYes)

R标准差以及开平方：sd(election$AnticipateYes) / sqrt(nrow(election))

Margin Of Error 误差范围：其实就是置信区间数值的一半。

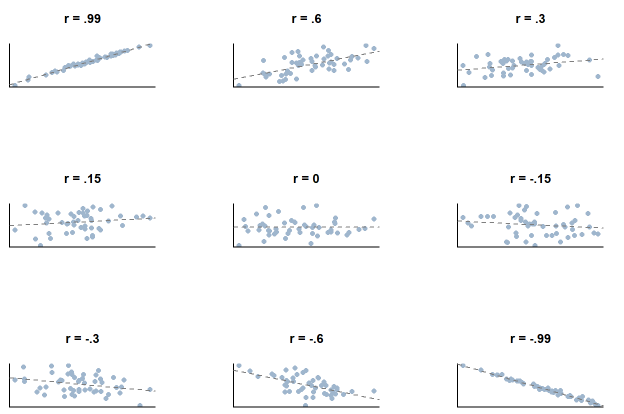


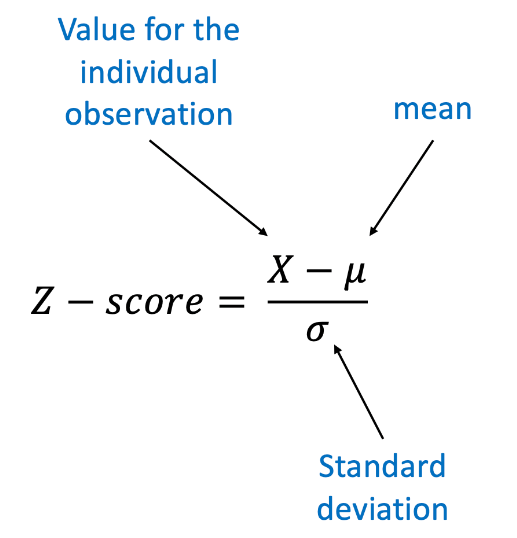
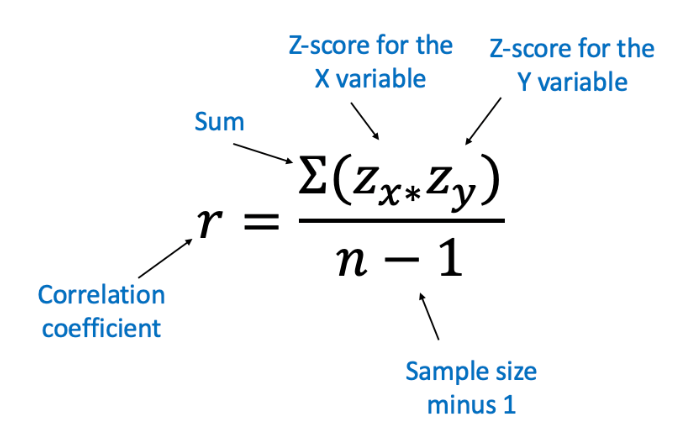
# Correlations

Correlation 相互关系：Correlation describes a relationship between two things based on patterns of change, where a change is one is asscoaited wtih a change in the other.

Correlation Coefficient 相关系数：一个数值或函数，用于表示两组数据或两个随机变量之间的相关程度，其值等于它们的协方差除以它们标准差的乘积。

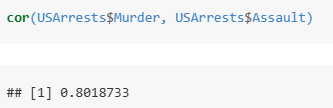
A correlation coefficient is a single number that can range from 1 to -1, and can fall anywhere in between.

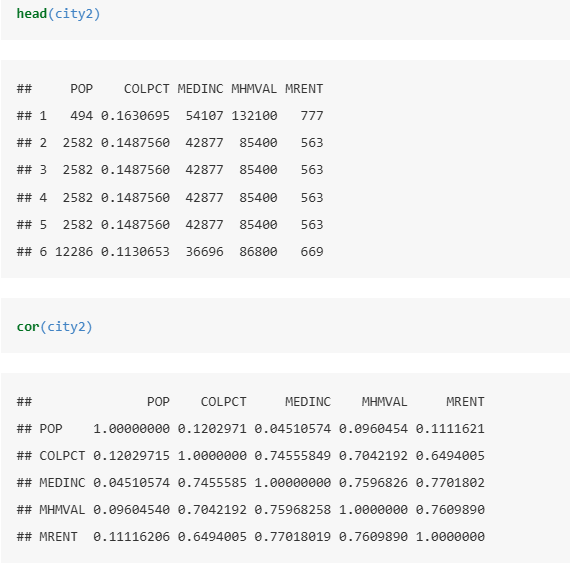


Strength of Correlation







Correlation and causation 相关性和因果关系：相关性指两个或多个变量之间的关系，而因果关系则指一个变量的变化导致了另一个变量的变化。相关性不一定意味着因果关系，因为可能存在其他未知的变量影响了两个变量之间的关系。

correlation is necessary for causation to be present, but it’s not sufficient on its own.

# Regression

regression 回归

Regression analysis at its most basic is fitting a line to our data to predict how changes in the x axis impact changes in the y axis.

##

## Call:

## lm(formula = wage ~ education, data = PSID1982)

##

## Residuals:

## Min 1Q Median 3Q Max

## -1038.6 -311.6 -48.1 222.3 3603.4

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 70.467 92.231 0.764 0.445

## education 83.888 7.017 11.955 <2e-16 \*\*\*

## ---

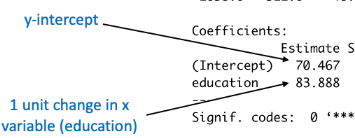
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

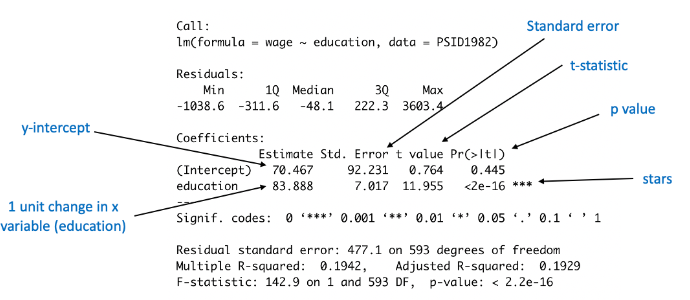
## Residual standard error: 477.1 on 593 degrees of freedom

## Multiple R-squared: 0.1942, Adjusted R-squared: 0.1929

## F-statistic: 142.9 on 1 and 593 DF, p-value: < 2.2e-16

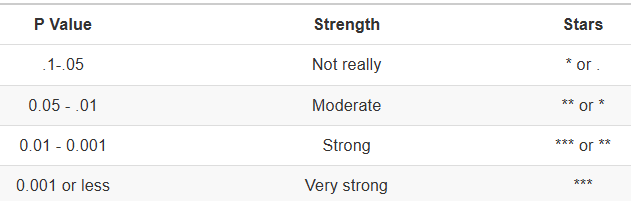


hypothesis testing 假设检验：一种统计学方法，用于验证关于总体参数的假设。它通过收集样本数据并进行统计分析来判断假设是否成立。



All the pieces are actually in the output above. We’re testing what the chances are that we’d get a coefficient as large as 83.888 for education if the actual relationship in 0. So to figure out the chances of that, we subtract our coefficient from our null relationship (0), and divide that figure by the standard error (which was already calculated for us). So 83.888-0 / 7.017, which equals 11.955. That is the t-statistic that is also in the output. Essentially, it’s asking the chances we’d get this large of a difference in the means depending on how tightly clustered our data is. And t statistics are similar to the z values we used in the chapter on polling. If we get a t statistic above 1.96, that means we will be more than 95% confident. Anyways, that’s all explained in more detail in another future chapter.

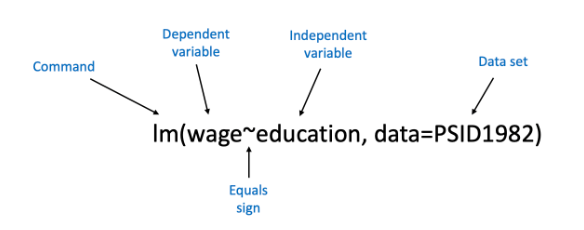


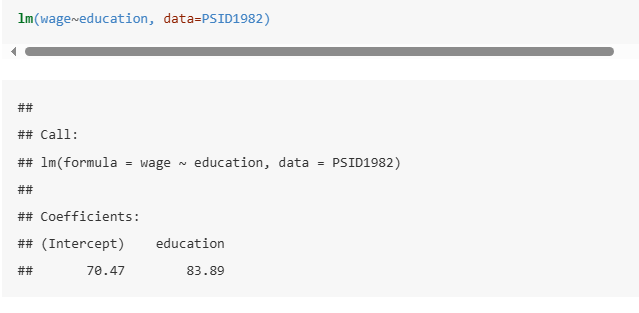


P value or Pr(>|t|)就是自变量不能影响因变量的概率。P value是由t value决定的，如上段所言，t value可以类比之前的Z值，即到Z为1.96时的信任度（Central Limit Theorem 中心极限定理）。

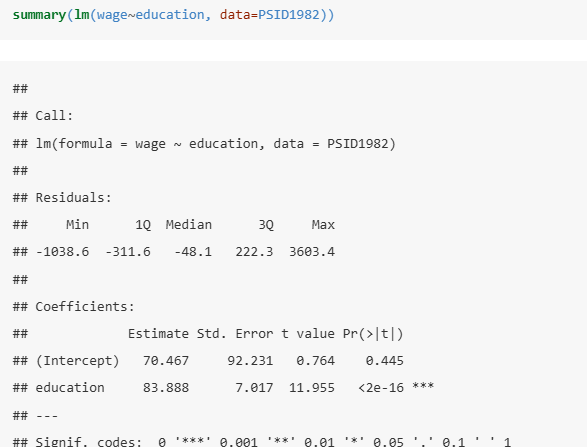
Practice

The basic command for running a regression is lm(). lm stands for linear model, which is what we’re estimating.





We get that additional output by asking for the summary of the regression model.



# Multiple Regression

多元回归：一种通过使用多个其他变量来估计一个变量的回归分析方法。

Multiple regression can help us try though. Multiple regression doesn’t mean running multiple regressions, it refers to including multiple variables in the same regression.

##

## Call:

## lm(formula = wage ~ ethnicity + occupation + education + experience +

## weeks + industry + south + smsa + married + gender + union,

## data = PSID1982)

##

## Residuals:

## Min 1Q Median 3Q Max

## -1008.8 -279.8 -37.4 196.7 3482.1

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) -39.735 234.478 -0.169 0.865491

## ethnicityafam -167.249 71.815 -2.329 0.020207 \*

## occupationblue -180.637 48.699 -3.709 0.000228 \*\*\*

## education 65.257 8.698 7.503 2.34e-13 \*\*\*

## experience 6.578 1.744 3.772 0.000179 \*\*\*

## weeks 2.962 3.529 0.839 0.401726

## industryyes 86.693 38.443 2.255 0.024496 \*

## southyes -59.662 40.791 -1.463 0.144110

## smsayes 169.833 38.978 4.357 1.56e-05 \*\*\*

## marriedyes 85.620 64.500 1.327 0.184883

## genderfemale -344.516 80.124 -4.300 2.00e-05 \*\*\*

## unionyes 25.782 41.761 0.617 0.537225

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 429.7 on 583 degrees of freedom

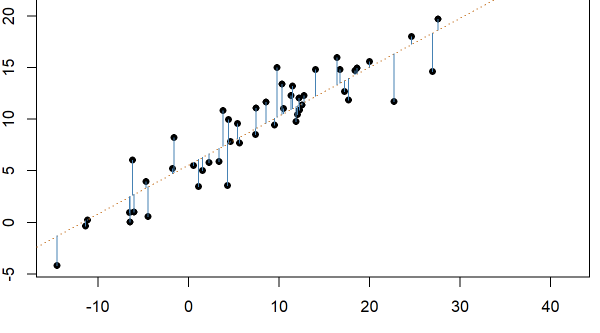
## Multiple R-squared: 0.3575, Adjusted R-squared: 0.3454

## F-statistic: 29.49 on 11 and 583 DF, p-value: < 2.2e-16

Yes, if you found two workers that had the same values for all of those variables except that they were of different races, the African American would still likely earn less.

# Evaluating Regression Models

ordinary least squares (OLS) 普通最小二乘法：回归分析中最常用的估计回归系数的方法，基于误差值之上。通过计算残差平方和（RSS）来估计回归系数。



in OLS the regression line is the line that achieves the least square criterion, whereby the sum of squared distances between the predicted and observed values are minimized.

. But in order for an OLS regression to be appropriate we have to assess whether **4 critical assumptions** are met.

The **four assumptions** are:

1. Linearity, the relationship between y and x is linear.
2. Independent, each observation is independent.
3. Homoscedasticity, the residuals have constant variance at every level. This is known as homoscedasticity.
4. Normality, the residual errors are normally distributed.

Assumption 1: Linearity

One typical way to test if the relationship between your dependent and independent variables is linear is by graphing the residuals from the model against the predicted values.

In general, the values should be **evenly** split above and below 0.

x <- rnorm(50, 10, 12) **#生成一个长度为50的向量，均值为10，标准差为12**

y <- (x\*.5)+rnorm(50, 5, 2)

x2 <- rnorm(60, 100, 40)

y2 <- (x2^2)+rnorm(60, 100, 500)

fit1 <- lm(y~x) **# 生成线性模型list，predict(fit)获得预测值**

fit2 <- lm(y2~x2) **# df1$predicted <- predict(fit)存储预测值到某dataframe**

res1 <- residuals(fit1) **#residuals残差：y值-预测值**

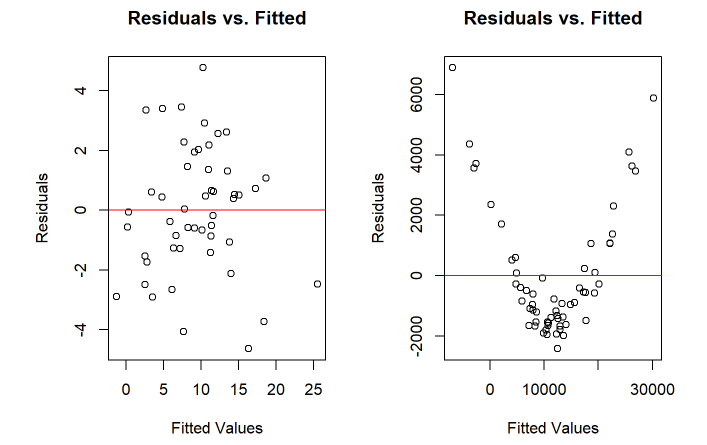
res2 <- residuals(fit2)

plot(predict(fit1), res1, xlab="Fitted Values", ylab="Residuals", main="Residuals vs. Fitted")

abline(0,0, col="red") **#添加横线？**

plot(fitted(fit2), res2, xlab="Fitted Values", ylab="Residuals", main="Residuals vs. Fitted") **#fitted应该等于predict**

abline(0,0, col="red")



Independence

Homoscedasticity 同方差性

概括而言，同方差说明回归的残差项是随机的，异方差说明残差项不随机，

举个例子：研究储蓄问题，收入低的人和收入高的人的储蓄其概率分布肯定不一样，想想穷学生和富大款，他们在决策储蓄这事上肯定不一样，当然富大款不一定储蓄就比穷学生多，可能还是负值，因为可以借债，但是穷学生借债就有限，形象点说，收入高的人的储蓄概率分布可能更广一点，因此方差比较大，如果穷学生和富大款都在同一组样本中，就会出现异方差。

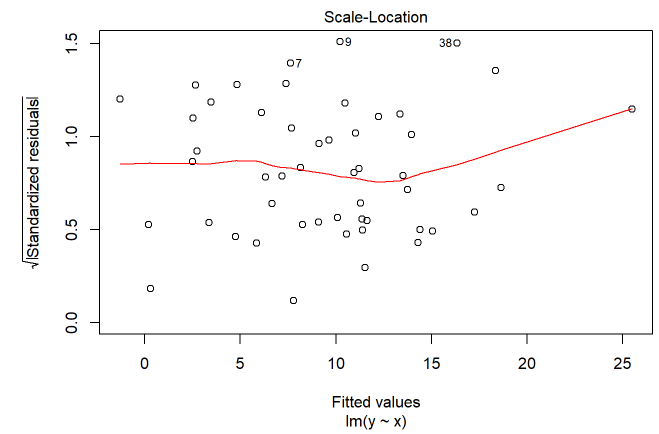
error term 残差项：我们不期望我们的回归模型能完美地预测y值，误差项捕获了因变量中不能由自变量解释的可变性的数量。

What would we need to do if there was heteroscedasticity in the model? Most likely, we’d need to add more independent variables to help explain.

library(car)

ncvTest(fit1)

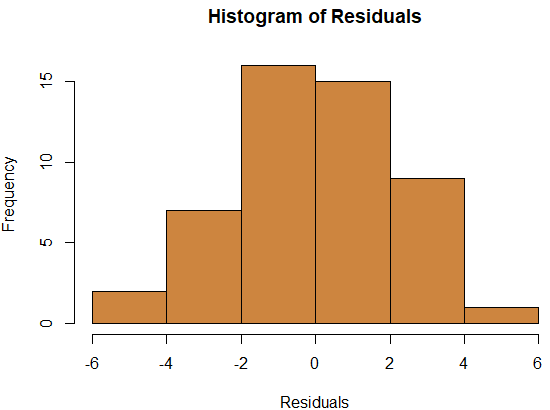
plot(fit1,3)



残差图分布比较随机，说明同方差性？

Normality： Our final assumption is that the residual errors from our model are normally distributed.

hist(res1, breaks=6, col="peru", main="Histogram of Residuals", xlab="Residuals")

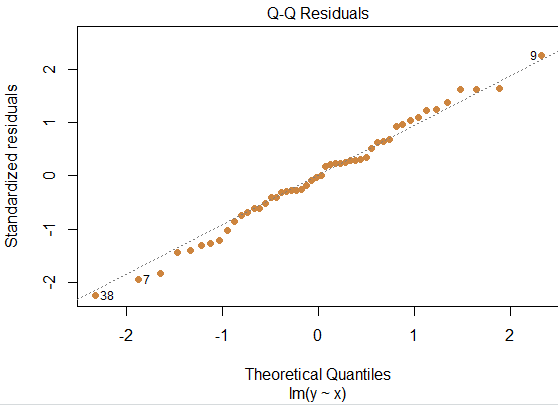


It’s not a perfect bell shape, but it’s close. And considering that we only have 50 observations in our data, it’s a pretty good approximation of evenly split above and below 0.

Q-Q Plot：it plots the quantiles for the fitted values of the regression against a standardized set of residuals.

what you want to see is a close match between the points and the 45 degree line. In the case below, we can again see the residuals are normal distributed.

plot(fit1, 2, col="peru", pch=16)



R-Squared

A prominent metric for evaluation your model is the R-Squared or R2. R-Squared can range from 0 to 1. and is commonly interpreted as indicating the share of the variance in the dependent variable that is explained by your model.

Higher values of R2 are preferred, because they indicate we’ve explained more of the variation in our dependent variable. We’ve done a better job of predicting our outcome.

同时注意R-Squared和adjusted R-Squared的区别，随着变量的增加，R-Squared值会无脑增加，后者则会由考虑的增加或减少，但通常后者会小于前者。

R-Squared is just a measure of how much variation the model explains, it says nothing about whether your model is correct.

R Squared不能用特定的标准来衡量，对于某些话题，如预测身高，R Squared通常会很高，因为身高比较好预测，通过如父母身高、体重、鞋子大小等；但是对于薪水等话题时，R Squared通常会比较低，因为薪水的预测远比身高复杂得多。因此一个模型的好坏不能仅依靠R Squared值来判断。如上段所言：他只是衡量模型解释偏差的程度。

Multicollinierity多重共线性：多重共线性是指独立的自变量之间具有高度的相关性，这意味着无法确定每个变量对因变量的唯一贡献。

variance inflation factors (VIF) 方差膨胀因子: 为了解多重共线性是否存在问题，我们需要在运行回归后计算方差膨胀因子(VIF)。高于5的VIF表明我们应该关注。

示例：vif(fit)

当vif值大于5时，有三个选择。

1. 你可以放弃你认为不太重要的那个

2. 您可以转换变量，例如创建一个新的变量

3. 如果你认为这两个变量在理论上对你的分析都很重要，你可以同时保留它们。

同时，多重共线性通常会放大变量的标准误差，从而降低显著性的可能性。因此，由于多重共线性，您不太可能说某个多重共线性变量很重要

如果所讨论的变量是你分析的核心，并且你认为它们是必要的，你可以同时保留它们。但是测试VIF总是值得的，看看是否有不必要的变量可能会影响您的结果。实际上，从上述回归中删除一个或两个变量不会对结果产生有意义的影响，这意味着无论我们选择哪个选项，它都不会产生实际的差异。最重要的是对它进行测试，并意识到潜在的问题，这样我们才能解释这个决定。

Corrr函数：除VIF外，为了验证非共线性，我们还可以使用corrr()包计算变量之间的积矩相关系数，这是tidymodels的一部分。在理想的情况下，我们应该寻找小于0.8的相关性

示例代码如下：

**library**(corrr)

Correlation <- LonWardProfiles **%>%**

**st\_drop\_geometry**()**%>%**

dplyr**::select**(average\_gcse\_capped\_point\_scores\_2014,

unauthorised\_absence\_in\_all\_schools\_percent\_2013,

median\_house\_price\_2014) **%>%**

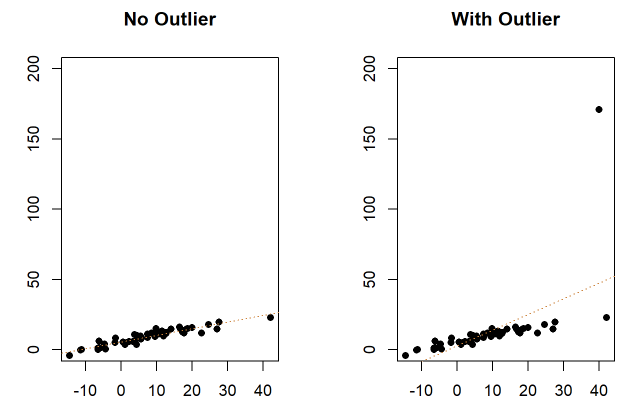
*# just focus on GCSE and house prices*

**focus**(**-**average\_gcse\_capped\_point\_scores\_2014, mirror = TRUE)

*#visualise the correlation matrix*

**rplot**(Correlation)

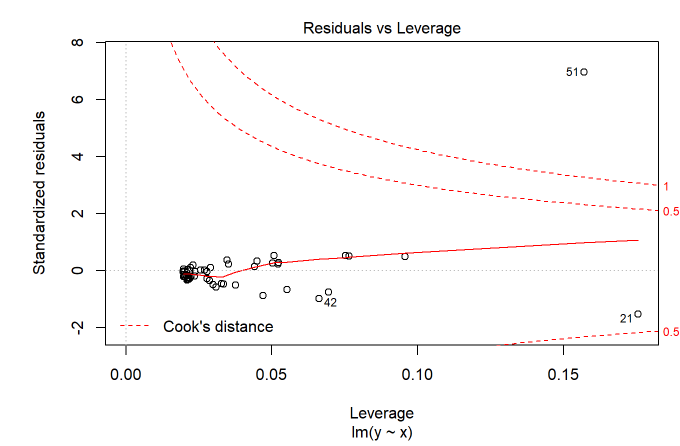
Outliers：Outliers are data points that don’t fit your model well.



Cook’s Distance: a measure of the influence or leverage of a data point.

库克D值意味着，如果我们放弃每个单独的观察结果，会对结果产生多大的影响。任何库克D值是平均值的3倍的值都被认为是大的。

下图显示了所有观测的库克距离。

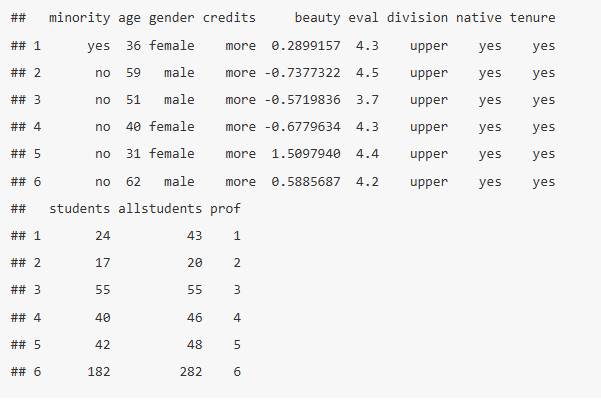


Practice

**首先导入数据**

TeachingRatings <- **read.csv**("https://raw.githubusercontent.com/ejvanholm/DataProjects/master/TeachingRatings.csv")

**head**(TeachingRatings)



**查看回归总结表**

**summary**(**lm**(eval~minority + age + gender + credits + beauty + division + native + tenure + students + allstudents, data=TeachingRatings))

**查看AIC赤池信息量准则，以判断是否保留tenure项。越小越好**

赤池信息量准则：它建立在熵的概念基础上，可以权衡所估计模型的复杂度和此模型拟合数据的优良性。

reg\_withtenure <- **lm**(eval~minority + age + gender + credits + beauty + division + native + tenure + students + allstudents, data=TeachingRatings)

reg\_notenure <- **lm**(eval~minority + age + gender + credits + beauty + division + native + students + allstudents, data=TeachingRatings)

**AIC**(reg\_withtenure)

## [1] 700.941

**AIC**(reg\_notenure)

## [1] 699.6271

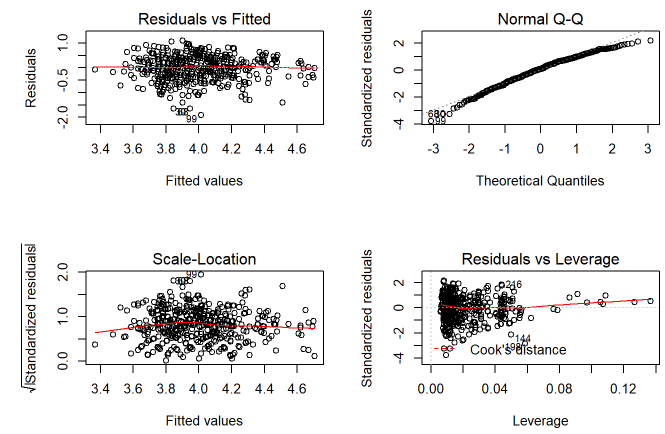
The regression without tenure as an independent variable had a lower AIC, so we would be justified in removing it.

**查看回归系列表**

通常会有4个总结回归结果的表。

**par**(mfrow=**c**(2,2)) #show in a 2 by 2 pattern

**plot**(reg\_notenure)



Residuals vs Fitted：检查线性回归模型是否合理。红线上下的点分布均匀即代表线性回归模型合理。x轴对应的是模型预测的值，y轴对应的是各预测值的残差（即各原值-预测值）

Normal Q-Q：用于检查残差是否符合正态分布。这些点需要尽可能的跟随那条直线。本质上y轴是排序数据，x轴是我们在数据确实来自正态分布时所期望的值。

Scale-Location：用于检查同方差性/检查残差中方差的同质性。我们想要看到的是中间有一条水平线上面和下面有相等的点;如果这条线不是水平的，那就表明数据的某些部分存在异方差。

Residuals vs Leverage：判断各点在被纳入或被排除分析时可能影响回归结果的情况。在图中以名称标识的任何行都被认为具有很大的杠杆作用，它们将被放置在虚线之外。这里，这些线之外没有点，也就是说没有异常值。如果有，它们可能需要额外的审查，但不建议您删除异常值。

**检查vif值（**方差膨胀因子）

**library**(car)

**vif**(reg\_notenure)

## minority age gender credits beauty division

## 1.225254 1.196876 1.136262 1.237300 1.141156 1.280655

## native students allstudents

## 1.142814 19.135157 19.526222

That makes sense, classes with more students are likely to have more evaluations submitted, even if a lower percentage do. We can make an adjustment.

TeachingRatings$evalpct <- TeachingRatings$students/TeachingRatings$allstudents

reg\_notenure2 <- **lm**(eval~minority + age + gender + credits + beauty + division + native + evalpct + allstudents, data=TeachingRatings)

**vif**(reg\_notenure2)

## minority age gender credits beauty division

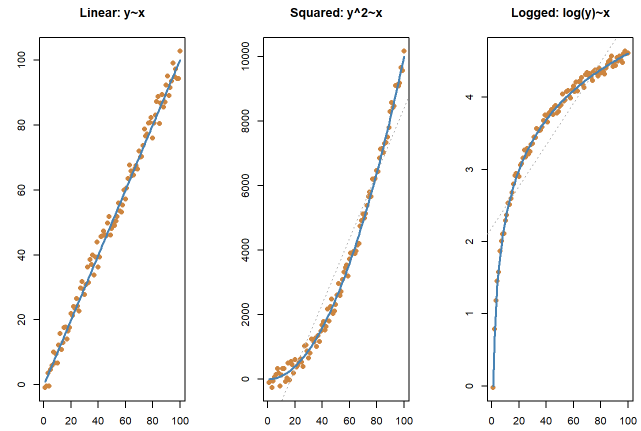
## 1.235526 1.199069 1.132739 1.235669 1.155962 1.248594

## native evalpct allstudents

## 1.144272 1.214753 1.269450

# Transforming Variables in Regression

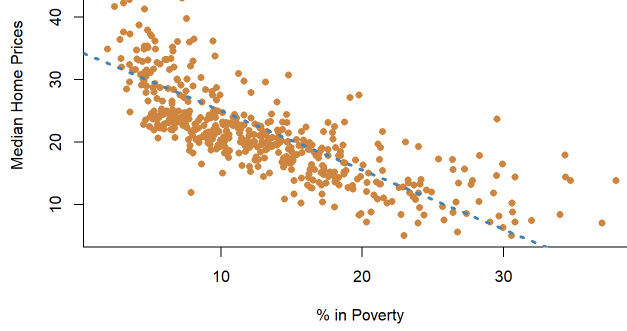
线性关系是自变量和因变量之间的一种关系，但它不是唯一的形式。在回归中，我们试图拟合一条最能代表预测器、自变量和因变量之间关系的线

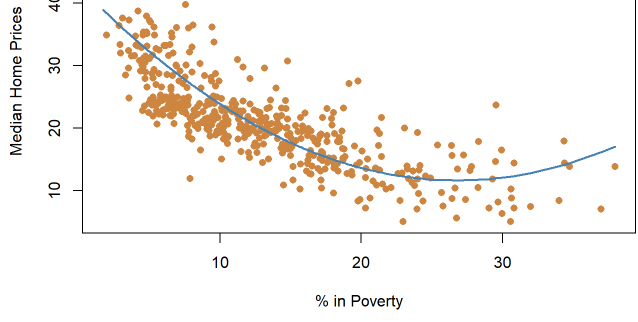


Polynomial Regressions 多项式回归：一种回归分析方法，通过拟合多项式函数来建立自变量和因变量之间的关系。

lm(formula = medv ~ lstat, data = Boston)

lm(formula = medv ~ lstat + lstat\_sq, data = Boston)





Logged variables 我们也可以考虑取因变量或自变量的对数。

## lm(formula = log(medv) ~ log(lstat), data = Boston)

//

PSID1982 <- read.csv("PSID1982.csv")

par(mfrow=c(2,2))

reg <- lm(wage~education,data=PSID1982)

plot(reg)

summary(reg)

AIC(reg)

"在残差与拟合图中，残差的均值看起来不像是0，而Q-Q图显示，高值偏离了直线。让我们

看看非线性变换是否能解决这些问题。我们先看看如果加上平方项会怎样。"

#要加上平方项，只需把变量名放在I()中，然后加上^2，表示你想要平方项。

par(mfrow=c(2,2))

reg2 <- (lm(wage~education+I(education^2), data=PSID1982))

plot(reg2)

summary(reg2)

AIC(reg2)

#That didn’t do a lot, although the AIC did go down slightly.

#Let’s see if using the log of wage improves the fit.

par(mfrow=c(2,2))

reg3 <- lm(log(wage)~education,data=PSID1982)

plot(reg3)

summary(reg3)

AIC(reg3)

#对工资这样的连续变量取对数是很常见的因为它们往往是非线性的。

"现在，让我们添加职业和教育之间的交互作用。我们可以添加交互作用，只需在我们感兴趣

的第二个变量中加上\*(乘法符号)，而不是加号。这助于我们理解，

对于白领和蓝领工人来说，接受额外教育的工资增长是否存在差异"

reg4 <- lm(log(wage)~education\*occupation, data=PSID1982)

summary(reg4)

AIC(reg4)

"When our dependent variable is logged we need to take the exponent of the

coefficient to interpret the size of change."

exp(0.04509) # exponent for the coefficient for education

## [1] 1.046122

"没有受过教育的白领工人的工资预计比没有受过教育的蓝领工人低21%，

但这种差异微不足道。这个结果本身并不完全有意义，因为我们期望白领工人赚得更多，

但这只是针对没有受过教育的工人，而数据中没有人真正受过教育。"

exp(-0.23689) # exponent for the coefficient for occupation

## [1] 0.7890781

"早些时候，我们发现教育提高了蓝领工人的工资。通过相互作用，我们发现，白领工人

的教育水平每提高一个单位，其工资增长就比蓝领工人的工资增长多2.7%，而且这种变化

略显重要。因此，所有工人都从教育中获得了积极的收益，但对白领工人来说，

工资的增长更大。"

exp(0.02701)# exponent for the coefficient for the interaction between education and occupation

## [1] 1.027378

//

# Spatial Models

Spatial autocorrelation 空间自相关：Spatial autocorrelation refers to how strongly objects correlate with other nearby objects across a spatial area.

We can also look for clustering using what’s called a Moran’s I test. Moran’s I tests for whether or not your residuals are clustered more than chance would predict; a significant p-value indicates that yes, the data is clustered.

莫兰指数p值如果小于0.05则说明存在聚集现象。

Neighbors 邻域：In order to address the spatial correlation for our residuals with either model we’ll need to tell R how each unit is clustered. 本质上，如果我们认为附近单位之间存在问题，我们需要首先确定哪些单位靠近其他单位。我们可以通过创建一个queens邻域矩阵来实现。

空间滞后模型、空间误差模型和空间杜宾模型都是空间计量经济学中常用的模型，主要用于分析空间数据的相关性和空间依赖性。

空间滞后模型是指在回归模型中加入空间滞后项，即因变量的值不仅受自变量的影响，还受到邻近地区因变量的影响。空间误差模型则是在回归模型中加入空间误差项，即因变量的误差项不仅与自变量的误差项相关，还与邻近地区的误差项相关。空间杜宾模型则是将空间滞后项和空间误差项同时加入回归模型中，综合考虑了空间滞后性和空间误差性的影响。

ols、空间滞后模型、空间误差模型、空间杜宾模型如何选取

OLS（普通最小二乘法）是一种广泛使用的经济学估计方法，它假设数据之间不存在空间相关性。因此，在选择模型时，如果我们认为数据之间存在空间相关性，OLS可能不是一个合适的选择。

空间滞后模型是一种考虑了空间相关性的模型，它假设因变量受到相邻地区的自变量影响。这种模型适合用于研究地区之间相互依赖的现象，例如城市之间互相影响的经济变量。当我们关注空间相关性时，空间滞后模型是一个合适的选择。

空间误差模型是另一种考虑了空间相关性的模型，它假设误差项与相邻地区的误差项存在空间相关性。这种模型适合用于研究个体观测值之间存在空间依赖的情况，例如地区的房价受到相邻地区房价的影响。当我们关注空间相关误差时，空间误差模型是一个合适的选择。

空间杜宾模型是一种将空间滞后模型和空间误差模型结合起来使用的模型，它同时考虑了因变量和误差项的空间相关性。这种模型适合用于研究地区间的相互影响，并且考虑了误差项的相关性。当我们关注同时存在因变量和误差项的空间相关性时，空间杜宾模型是一个合适的选择。

lm检验 空间滞后 空间误差 空间杜宾

lm检验（Lagrange Multiplier test）是一种常用的统计检验方法，用于检验模型的异方差性或共线性问题。它基于最大似然方法，通过构建一个特定的Lagrange乘子统计量来进行假设检验。通过计算Lagrange乘子统计量的值与相应的临界值进行比较，就可以判断模型是否存在异方差性或共线性问题。

空间滞后（Spatial Lag）指的是在空间经济模型中，考虑到邻近地区之间的相互影响，选择变量的值不仅与自身特征有关，还与其邻近地区的特征有关。空间滞后模型使用一种权重矩阵来反映地理邻近性，通过将自变量的值与权重矩阵相乘，建立了变量之间的空间相关关系，从而更准确地描述了数据的相关性。

空间误差（Spatial Error）是空间经济模型中的一种特殊形式，它假设误差项存在空间自相关性。与传统误差项独立同分布的假设不同，空间误差模型考虑了邻近地区之间的相关性，通过引入一个空间权重矩阵来表示地理邻近性。这样可以在模型中捕捉到由于空间相关性导致的相关误差项。

空间杜宾（Spatial Durbin）模型是空间经济学领域中的一种常用模型，它综合了空间滞后模型和空间误差模型的优点，能够处理包括空间滞后项和空间误差项的空间相依性。空间杜宾模型是在传统经济计量模型的基础上，引入了空间滞后和空间误差项，通过加入这两项来捕捉邻近地区之间的空间关联关系，提高模型的解释能力和预测精度。

Likelihood Ratio, Wald和Lagrange Multiplier检验的区别与相似点

似然比检验(Likelihood Ratio Tests/LR)：被用来评估两个模型并且比较两个模型的拟合效果。从一个模型中删除掉几个预测变量往往会使模型拟合效果变差，但这对于检验所观察的模型拟合度是否具有统计显著性来说是必要的。LR通过这种方式来比较两个模型的对数似然值来检验两个模型， 如果此差异（两个模型的对数似然值差异）是统计显著的，那么限制性更小的模型（参数更多的模型）相对限制性更大的模型对数据的拟合更好。

Wald检验：Wald与LR相似，但比LR要简单，因为它只需要评估一个模型。Wald通过检验的工作原理是检验一组参数等于某个值的零假设。对被检测的模型来说， 零假设是指感兴趣的两个系数是否同时为零。 如果检验结果无法拒绝零假设， 表明移除这两个变量将不会严重影响模型对数据的拟合效果， 因为相对系数标准差很小的系数通常对因变量的预测没有太大帮助。

Lagrange Multiplier检验：与Wald检验一样，Lagrange乘数检验仅需要估计一个模型。区别在于，使用拉格朗日乘数检验时，估计的模型不包含感兴趣的参数。

# Spatial Regression Analysis

空间权重 Spatial Weights：描述在空间分析中确定“邻域”的各种方法。我们将研究两种常见的方法:相邻性Contiguity(公共边界)和距离(k近邻)。

相邻性Contiguity：有Queens Case、Rooks Case、Bishops Case。

距离：Neighbors can also be assigned by distance or K-nearest neighbors. Using this method, distance is measured from the centroid of a polygon until a specific number of neighbors are achieved.

