**Presentation 3.64 ~ 3.67 Tips**

**A.goal:(见书3.4，分析内容详解)**

1.

Is there a relationship between ... and ...?

Determing whether the data provide evidence of an association between ... and ...

evidence weak or strong.

2.

How strong is the relationship between ... and ...?

giving X-data predict Y-data with a high level accuracy.

3.

If there are many qualities of dataset, which quality contribute to Y?

Find the function to saparate out the individual effects.

4.

How accurately can we estimate the effect of each quality.

5.

How accurately can we predict? Compare the predict-value and true-value.

6.

relationship linear or not?

If leaner, we use linear regression approach.

If not leaner, try to transform the predictor or the response so that linear regression can be used.

7.

synergy effect or interraction effect协同效应 can we find?

**B.Need find:**

1.intercept

2.slope

3.other coefficient/parameter

Approach: least squares

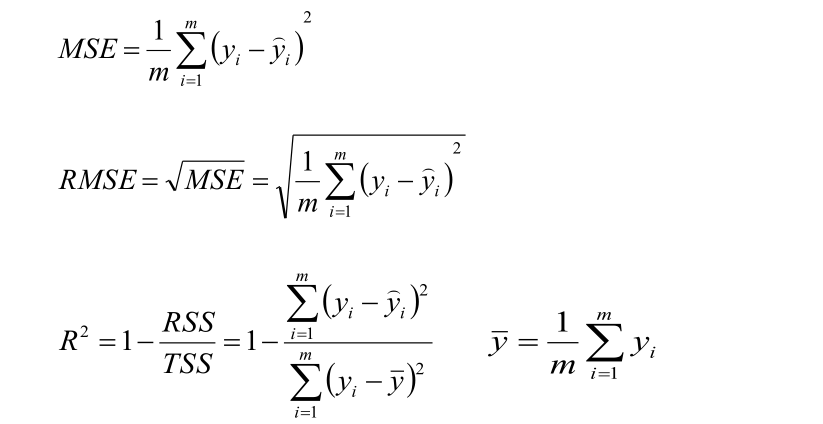
**C.Work**

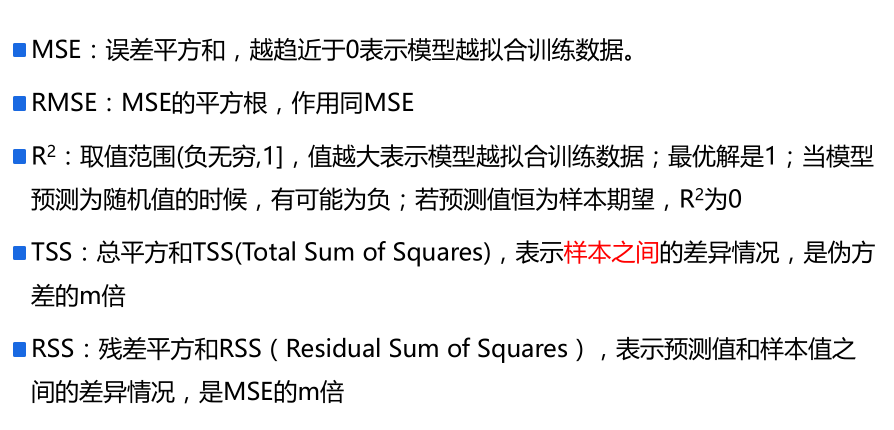
R-code, analysing, figures, PPT, tips, questions, speech.

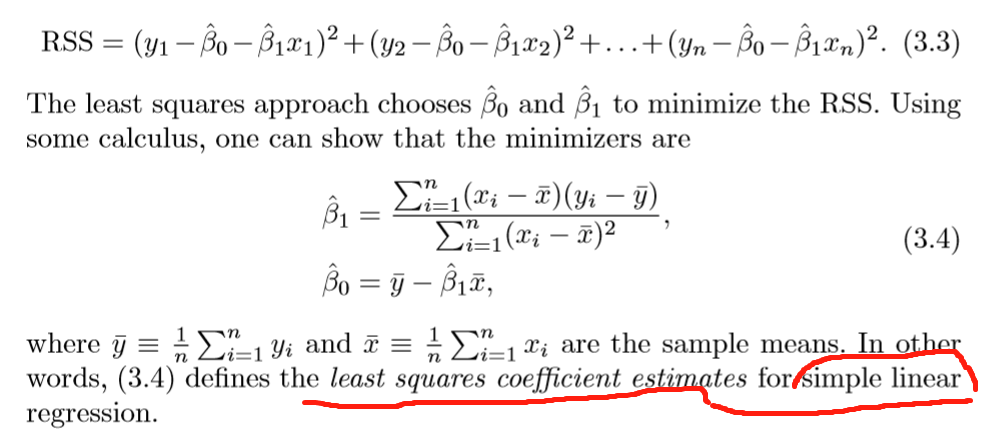
**D.Optional**

Tensorflow verification

**第三章所有概念:**





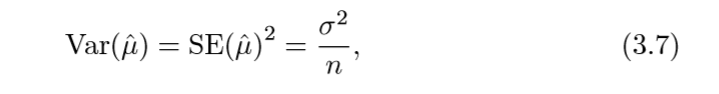


**population regression line**

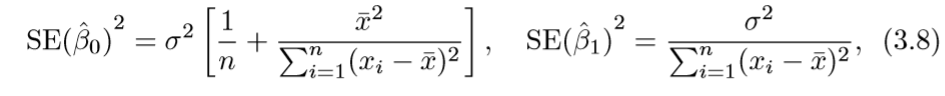


Unbiased and biased 参见6865，这里用unbiased

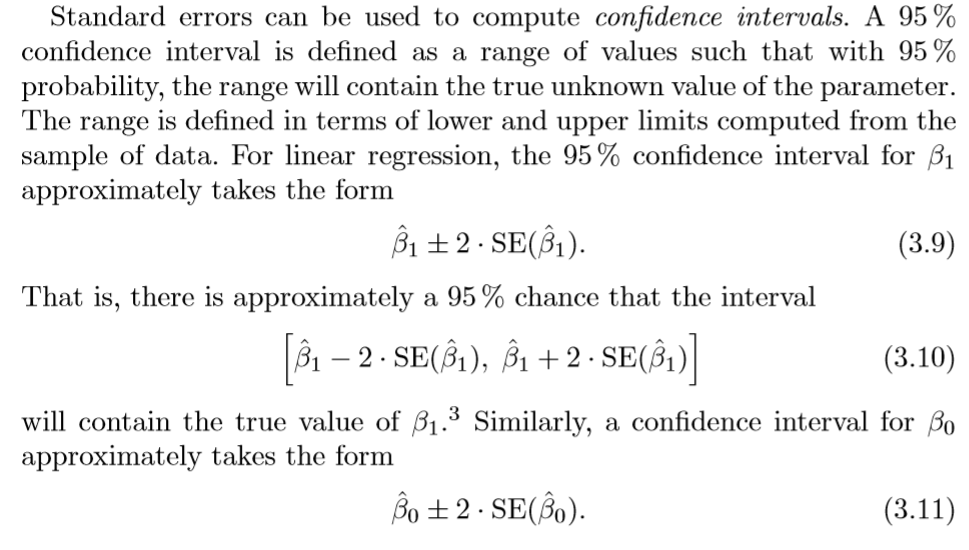
**Estimate how far oﬀ will that single estimate of ˆ μ be**

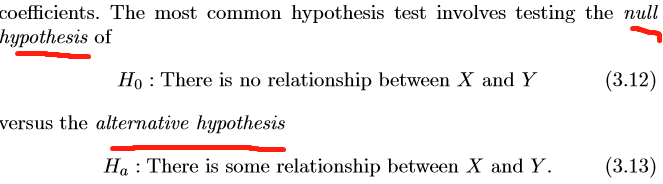


To compute the standard errors associated with ˆ β0 and ˆ β1, we use the following formulas:



置信区间：confidence intervals.





**t-statistic**



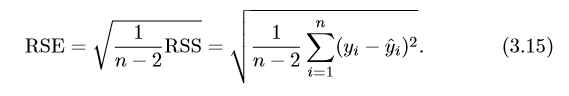


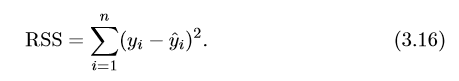
**p-value**

it is a simple matter to compute the probability of observing any number equal to |t| or larger in absolute value, assuming β1= 0.

if we see a small p-value,Roughly speaking, we interpret the p-value as follows: a small p-value indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance, in the absence of any real association between the predictor and the response. We reject the null hypothesis—that is, we declare a relationship to exist between X and Y —if the p-value is small enough. Typical p-value cutoﬀs for rejecting the null hypothesis are 5 or 1%

**Residual Standard Error**



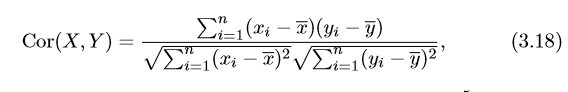


**R2 Statistic**



**R2 measures the proportion of variability in Y that can be explained using X. (0,1)**

**Correlation**

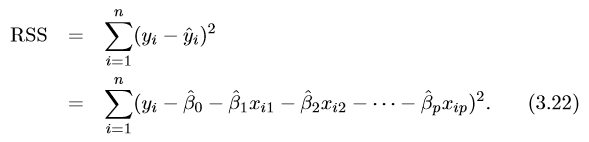
 

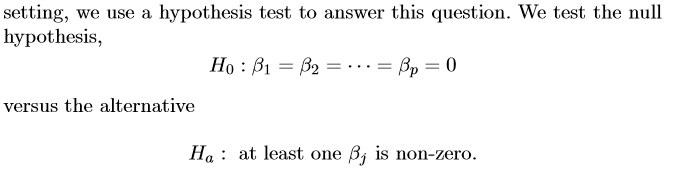
**相关性分析很重要，参考报纸对sales贡献负数，但报纸与广播却有0.35相关性的例子。鲨鱼袭击与冰激凌销售例子（隐藏因子：温度）**

**Multiple Linear Regression**

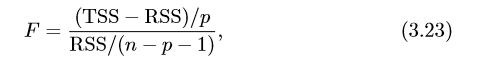


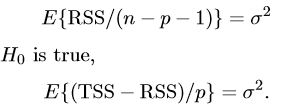






**F-statistic**

**即当F =1 时H0成立，F与n和p的值有关，当n很大p很小时，at least one of quality相关Ha。变量Variable and 响应response 关联associated.**

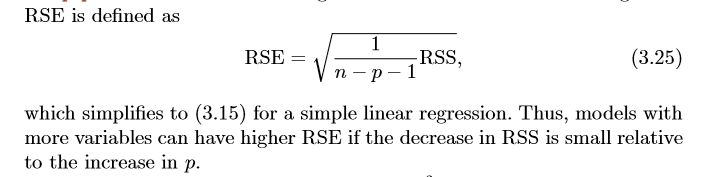
**F-statistic 可以在t=statistic和p值可能错误断定时提供进一步的检验，if H0 is true, there is only a 5% chance that the Fstatistic will result in a p-value below 0.05**

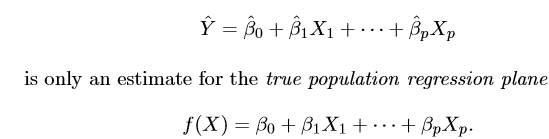
**当p>n时，目前F-statistic等本章各种判定不能使用。但第六章前向选择可以使用。**

**多元回归，优先计算F-statistic与p-value**

**P很大情况p>n，前向选择，反向选择，混合选择。**

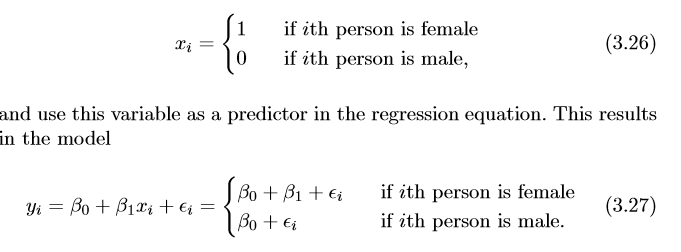
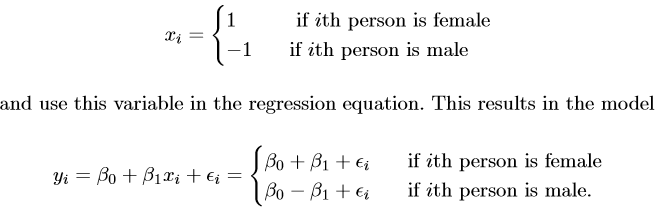
**3.15 RSE是简化的**



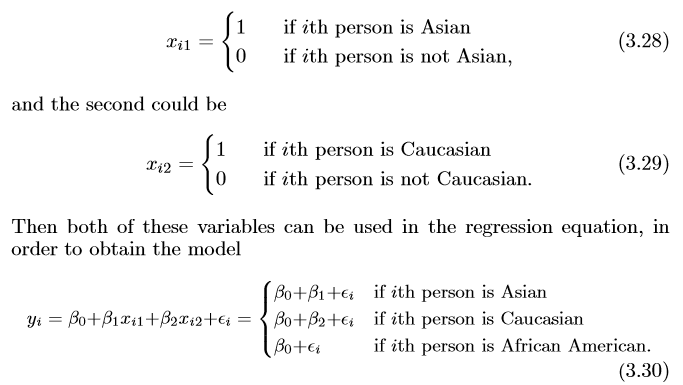


**Qualitative Predictors**

**编码方式不同，预测是一样的，创建虚拟变量dummy variable**

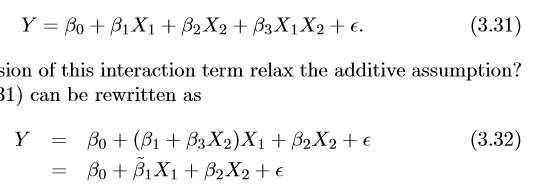
 

**对于多个levels，需要多个dummy variable**



**Additive assumption，包含交互项的模型，注意hierarchical principle**

**.A small p-value associated with the interaction term indicates the presence of such relationships.**

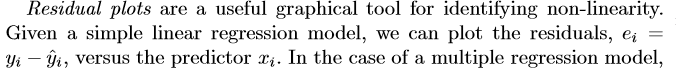


**polynomial regression（第七章）**

**very simple way to directly extend the linear model to accommodate non-linear relationships**

**添加多次项拟合非直曲线。**

**residual plot**

**for identifying none-linearity 大多数情况下判断是否非线性****见书P93**

**Correlation of Error Terms 如果误差不相关，P=value会降低导致错误估计。**

**In general, the assumption of uncorrelated errors is extremely important for linear regression as well as for other statistical methods, and good experimental design is crucial in order to mitigate the risk of such correlations.书P95**

**Non-constant Variance of Error Terms书P96**

**Heterosceda sticity. 异质性**

**outer quantiles of the residuals.**

**Weighted least squares**



**Outliers 异常值**

**An outlier is a point for which yi is far from the value predicted by the outlier model.**

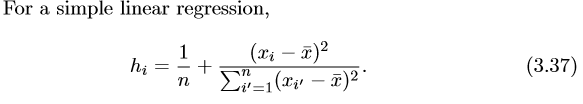
**instead of plotting the residuals, we can plot the studentized residuals, computed by dividing each residual ei by its estimated standard studentized residual error.**

**one solution is to simply remove the observation.**

**High Leverage Points**

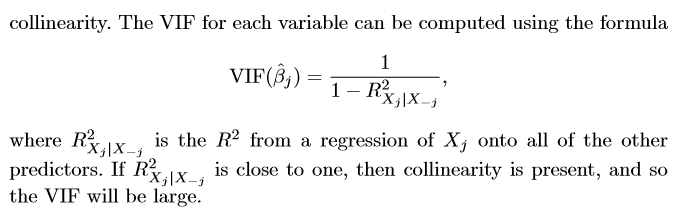
**removing the high leverage observation has a much more substantial impact on the least squares line than removing the outlier. it is important to identify high leverage observations.**

**Leverage statistic 计算高杠杆点**



**if a given observation has a leverage statistic that greatly exceeds (p+1)/n, then we may suspect that the corresponding point has high leverage.**

**Collinearity 共线性**

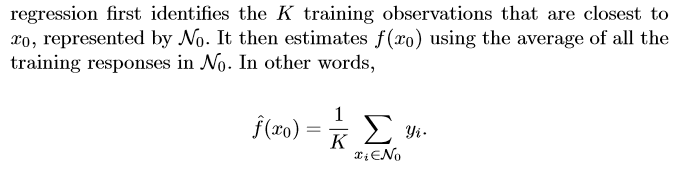
**it causes the standard error for ˆ βj to grow. VIF 值一般超过5或者10就会有共线性。**

**解决方法：**

**The ﬁrst is to drop one of the problematic variables from the regression.**

**The second solution is to combine the collinear variables together into a single predictor.**

**Comparison of Linear Regression with K-Nearest Neighbors（KNN）**



**As a general rule, parametric methods will tend to outperform non-parametric approaches when there is a small number of observations per predictor.**

**P很大导致curse of dimensionality ，使得KNN变差**