第三部分：回归

我们对这个数据集以粉丝数作为响应变量做回归，观察到粉丝数全是非负的整数，属于计数数据，具有厚尾分布（图），右偏大约8.56。对于计数数据，一种方法是把它当作连续性变量，做log变换后，进行多元线性回归；或者去拟合poisson回归或负二项回归。

1.数据预处理：

鉴于self\_tags标签五花八门，方便起见，我们只关心该视频是否有标签，将self\_tags转为0，1类别变量；

我们将离散变量sex, video\_tag\_combine，self\_tags 设为factor；

综合与第三方平台信息，简单叠加微博、微信、qq等等各列，得到新变量关联平台数；

对粉丝数、发表视频数、播放量等等右偏的变量进行log变换以移除skewness。

2.多元线性回归

基于变换后的数据，我们对这些连续型的预测变量关于粉丝数绘制散点图（图），拟合直线，结合之前变量选择得到的结果（图），发现视频播放量，关联平台数、优秀作品数、发表视频数、发表相册数与粉丝数间都具有明显线性关系。特别的，这里显示视频时长对粉丝数有负向线性作用。

初步拟合：

我们对这五个变量视频播放量，关联平台数、优秀作品数、发表视频数、发表相册数外加视频标签这一类别变量进行初步拟合（结果图），发现所有变量p值均显著，R方达到0.72，AIC值为23265，F检验的p值也非常显著，说明我们模型是有用的。

StepBIC：

接下来进行stepBIC，新加入视频时长和完播率两个变量，这两个变量也在之前变量选择结果中出现过，对这8个变量重新拟合（结果图），此时R方提升了一些，将近0.73，AIC值为23086。

共线性；

值得注意的是，这里视频时长拟合的系数为正，但视频时长与粉丝数是负相关的，说明有可能发生了共线性，我们计算vif（图），发现所有变量的vif都小于10，说明没有共线性。

模型诊断：

我们对模型的假设进行验证：

1. 正态性，Q-Q plot（图）画出来还可以，正态性满足；
2. Observations独立性，进行durbinWatsonTest（图）发现p值不显著，说明观测值间没有autocorrelation；
3. 画crplots（图）检验线性性，可以看到红线与蓝线几乎重合，线性性也满足的很好；
4. 方差齐性，从残差图（图）可以看到，方差先增大再减小，做ncvTest（图）也发现，p值非常小，说明出现异方差。

解决异方差：

一种方法是使用广义最小二乘法（图 OLS,GLS,WLS），这个是史老师上课讲到的，假设s 的方差是sigma方V，而不是之前的sigma方 identity，然后最小化这个东西，拟合得到AIC值为22880，相比之前又降低了一些，拟合结果见图。

还有种常见方法是使用加权最小二乘法，相比较于一般的OLS，WLS就是在这前面加了个权重，这里wights采用上一模型拟合残差的平方 分之一，拟合后的R方是0.84（图），有了很显著的提高。且此时视频时长前的系数为负，符合原先负相关的关系。画出加权残差图（图）后发现，异方差问题得已解决。

3.广义线性模型

对于计数数据，一般经常做poisson回归和负二项回归，但我们计算粉丝数的均值和方差，均值为， 方差为，发现其均值远远小于其方差，而泊松回归是假设均值与方差相等，说明可能存在过度离散，我们用qcc包中的qcc.overdispersion.test进行检验，p值为0，发现确实存在overdispersion，所以我们进行负二项回归。

负二项回归：

负二项回归中其均值是u，方差是u+alpha u方，alpha等于1比theta，alpha为0时，就是泊松回归。它和泊松回归具有同样的均值结构，就是假设log(u)=Xbeta，即log(u)是预测变量的一些线性组合，只是多了一个参数alpha去表示数据的离散程度，估计的dispersion参数对粉丝数的均值没有影响，只是影响了方差。

拟合后的theta值为1.29，1比theta，也就是alpha不为0，为0的时候就是泊松回归，这也说明我们选用负二项回归是合理的。但是模型AIC值远高于前面GLS的AIC，说明负二项回归拟合效果不是很好，这可能由于我们没有加入时间轴，如果考虑到各个up主入站的时间，负二项回归拟合效果应该会更好。

模型解释：

考虑GLS、WLS和负二项回归，三个模型结构非常相似，GLS和WLS模型是把粉丝数看作连续变量，做log变换移除skewness；负二项回归则是假设log(u)是预测变量的线性组合。

由于视频播放量与粉丝数相关性最大，从拟合的回归模型，我们定量分析视频播放量的变化对粉丝数的影响，当视频播放量增大100，约增加16个粉丝；视频播放量增大1000，约增加63个粉丝。对于分类变量，三个模型都显示，鬼畜视频相对不吸粉，时尚视频相对吸粉。