Regression Analysis

Multiple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

Predicting Demand for Rental Bikes: P-values and Large Sample Size

Georgia Tech

1

About This Lesson

Georgia Tech

2

The P-value Problem: Basis Statistics

- Basic statistics under large sample size:

$Z_1, \ldots, Z_n \sim N(\mu, \sigma^2) \Rightarrow \bar{Z} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

- Hypothesis testing for the mean:

$H_0\colon \mu = 0$ vs. $H_A\colon \mu \neq 0$

- P-value and sample size:

$p - value = 2P(Z > \sqrt{n}|\frac{\bar{Z}-0}{\sigma}|)$ is approximately *0* with *n* very large

"Inflated" Significance: Conclusions based on small-sample statistical inferences using large samples can be misleading.

Samples Can Make the Insignificant...Significant!

Georgia Tech

3

The P-value Problem: Regression Analysis

- Hypothesis testing for the statistical significance of the regression coefficients:

$H_0\colon \beta_i = 0$ vs. $H_A\colon \beta_i \neq 0$

- P-value and sample size:

$p - value = 2P(T_{n-p-1} > |t - value|)$ is approximately *0* with *n* very large

- Misleadingly, reject the null hypothesis of zero coefficient – all or most relationship are statistically significant.

"Inflated" Statistical Significance: Conclusions based on small-sample statistical inferences on the regression coefficients using large samples can be misleading.

Georgia Tech

4

The P-value Problem: Approach

- Sub-sampling: Sample the observed data, e.g. 10-20% of the sample size
- Apply the regression model to each sub-sampled data
- Repeat for B times, e.g. B=100
- **Output**:

  Sub-sample 1: $\hat{\beta}_{0,1}, \hat{\beta}_{1,1}, \dots, \hat{\beta}_{p,1}$ & corresponding p-values $pv_{0,1}, pv_{1,1}, \dots, pv_{p,1}$

  Sub-sample 2: $\hat{\beta}_{0,2}, \hat{\beta}_{1,2}, \dots, \hat{\beta}_{p,2}$ & corresponding p-values $pv_{0,,2}, pv_{1,2}, \dots, pv_{p,2}$

  .....

  Sub-sample B: $\hat{\beta}_{0,B}, \hat{\beta}_{1,B}, \dots, \hat{\beta}_{p,B}$ & corresponding p-values $pv_{0,B}, pv_{1,B}, \dots, pv_{p,B}$
- Empirical distributions of the regression coefficients and the p-values

Georgia Tech

5

The P-value Problem: Approach

- Sub-sampling: Sample the observed data, e.g. 10-20% of the sample size
- Apply the regression model to each sub-sampled data
- Repeat for B times, e.g. B=100
- **Output**:

  Sub-sample 1: $\hat{\beta}_{0,1}, \hat{\beta}_{1,1}, \dots, \hat{\beta}_{p,1}$ & corresponding p-values $pv_{0,1}, pv_{1,1}, \dots, pv_{p,1}$

  Sub-sample 2: $\hat{\beta}_{0,2}, \hat{\beta}_{1,2}, \dots, \hat{\beta}_{p,2}$ & corresponding p-values $pv_{0,,2}, pv_{1,2}, \dots, pv_{p,2}$

  .....

  Sub-sample B: $\hat{\beta}_{0,B}, \hat{\beta}_{1,B}, \dots, \hat{\beta}_{p,B}$ & corresponding p-values $pv_{0,B}, pv_{1,B}, \dots, pv_{p,B}$
- Empirical distributions of the regression coefficients and the p-values

Theoretical Underpinning:
- Statistical significance (or lack of it) can be identified based on the distribution of the p-values; specifically, if the empirical distribution is approximately uniform between 0 and 1, then we don't have statistical significance.
- Statistical significance (or lack of it) can be identified based on the confidence interval of the regression coefficient derived from the empirical distribution.

Georgia Tech

6

The P-value Problem: Approach (cont'd)

```
## Approach: Subsample 40% of the initial data sample & repeat 100 times
count = 1
n = nrow(train)
B = 100
ncoef = dim(summary(model1)$coeff)[1]
pv_matrix = matrix(0,nrow = ncoef,ncol = B)
while (count <= B) {
    # 40% random sample of indices
    subsample = sample(n, floor(n*0.4), replace=FALSE)
    # Extract the random subsample data
    subdata = train[subsample,]
    # Fit the regression for each subsample
    submod = lm(sqrt(cnt)~.,data=subdata)
    # Save the p-values
    pv_matrix[,count] = summary(submod)$coeff[,4]
    # Increment to the next subsample
    count = count + 1
}
# Count pvalues smaller than 0.01 across the 100 (sub)models
alpha = 0.01
pv_significant = rowSums(pv_matrix < alpha)
```

Georgia Tech

7

Statistical Significance

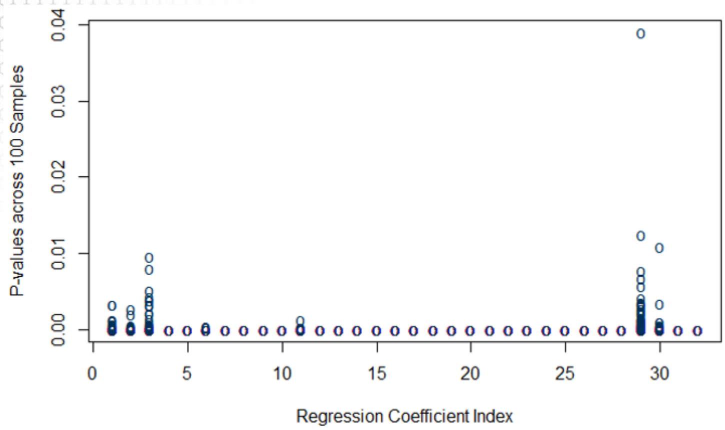
```
## Which regression coefficients are statistically significant?
idx_scoef = which(pv_significant>=95)
## Show the p-values of the significant coefficients in model2
cbind(summary(model2)$coeff[idx_scoef,c(1,4)],
     Freq=pv_significant[idx_scoef])
## Plot the 100 p-values of the significant coefficients
matplot(pv_matrix[idx_scoef,],
    xlab="Regression Coefficient Index",
    ylab="P-values across 100 Samples",
    type="p",
    pch="o",
    col=gtblue)
```

	Estimate	Pr(>|t|)	Freq
(Intercept)	1.670	0	100
season2	1.370	0	100
season3	1.380	0	100
season4	2.720	0	100
yr1	2.800	0	100
hr1	-1.630	0	100
hr2	-2.570	0	100
hr3	-3.770	0	100
hr4	-4.190	0	100
hr5	-2.360	0	100
hr6	1.480	0	100
hr7	6.820	0	100
hr8	10.700	0	100
hr9	7.500	0	100
hr10	5.440	0	100
hr11	6.210	0	100
hr12	7.450	0	100
hr13	7.310	0	100
hr14	6.770	0	100
hr15	7.090	0	100
hr16	9.020	0	100
hr17	12.700	0	100
hr18	12.100	0	100
hr19	9.440	0	100
hr20	7.020	0	100
hr21	5.380	0	100
hr22	3.860	0	100
hr23	2.000	0	100
holiday1	-0.986	0	98
weekday5	0.723	0	99
weathersit3	-2.650	0	100
hum	-2.580	0	100

Georgia Tech

8

Statistical Significance (cont'd)



Statistical significance: Most P-values are small across all sub-samples

Georgia Tech

9

Lack Statistical Significance

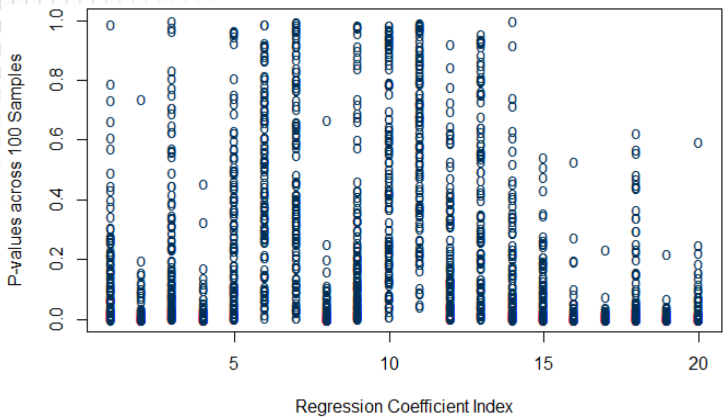
```
# Which regression coefficients are not statistically significant?
idx_icoef = which(pv_significant<85)
# Show the p-values of the significant coefficients in model2
cbind(summary(model2)$coeff[idx_icoef,c(1,4)],
      Freq=pv_significant[idx_icoef])
# Plot the 100 p-values of the significant coefficients
matplot(pv_matrix[idx_icoef,],
       xlab="Regression Coefficient Index",
       ylab="P-values across 100 Samples",
       type="p",
       pch="o",
       col=gtblue)
```

	Estimate	Pr(>|t|)	Freq
mnth2	0.379	0.005	12
mnth3	0.676	0.000	68
mnth4	0.516	0.021	11
mnth5	1.108	0.000	66
mnth6	0.499	0.043	7
mnth7	-0.326	0.240	1
mnth8	0.300	0.267	2
mnth9	1.052	0.000	64
mnth10	0.516	0.020	7
mnth11	-0.241	0.260	1
mnth12	-0.038	0.826	0
weekday1	0.229	0.024	9
weekday2	0.174	0.080	4
weekday3	0.283	0.004	16
weekday4	0.344	0.001	35
weekday6	0.530	0.000	79
weathersit2	-0.346	0.000	74
temp	3.847	0.000	38
atemp	4.879	0.000	84
windspeed	-1.101	0.000	59

Georgia Tech

10

Lack Statistical Significance



Lack of statistical significance: Uniform Distribution of P-values

11

Statistical Significance Summary

- Most regression coefficients remain statistically significant for 95% of the sub-samples, supporting statistical significance for these factors
- Statistical significance is not supported for most of months and weekdays as well as for temperature and windspeed factors given that other relevant factors, such as season and weather situation are in the model.
- While the 85% cutoff was used for the frequency of p-values being smaller than the significance level 0.01, other lower cut-offs, such as 50%, can be used.

12

Summary

13