Regression Analysis

Multiple Linear Regression

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

Ranking States by SAT Performance: Regression Analysis

1

About This Lesson

2

Linear Regression Analysis in R

```
regression.line = lm(sat ~ takers + rank + income + years + public + expend)
summary(regression.line)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-94.659109	211.509584	-0.448	0.656731	
takers	-0.480080	0.693711	-0.692	0.492628	
rank	8.476217	2.107807	4.021	0.000230	***
income	-0.008195	0.152358	-0.054	0.957353	
years	22.610082	6.314577	3.581	0.000866	***
public	-0.464152	0.579104	-0.802	0.427249	
expend	2.212005	0.845972	2.615	0.012263	*

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

Residual standard error: 26.34 on 43 degrees of freedom
Multiple R-squared: 0.8787, Adjusted R-squared: 0.8618
F-statistic: 51.91 on 6 and 43 DF, p-value: < 2.2e-16

Georgia Tech

3

Linear Regression Analysis in R

```
regression.line = lm(sat ~ takers + rank + income + years + public + expend)
summary(regression.line)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-94.659109	211.509584	-0.448	0.656731	
takers	-0.480080	0.693711	-0.692	0.492628	
rank	8.476217	2.107807	4.021	0.000230	***
income	-0.008195	0.152358	-0.054	0.957353	
years	22.610082	6.314577	3.581	0.000866	***
public	-0.464152	0.579104	-0.802	0.427249	
expend	2.212005	0.845972	2.615	0.012263	*

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

Residual standard error: 26.34 on 43 degrees of freedom
Multiple R-squared: 0.8787, Adjusted R-squared: 0.8618
F-statistic: 51.91 on 6 and 43 DF, p-value: < 2.2e-16

Test for statistical significance:

$\hat{\beta}_{takers}$ Pr(>|t|) ≈ 0.4926 > 0.1
$\hat{\beta}_{rank}$ Pr(>|t|) ≈ 0.0002 < 0.1
$\hat{\beta}_{income}$ Pr(>|t|) ≈ 0.9574 > 0.1
$\hat{\beta}_{years}$ Pr(>|t|) ≈ 0.0009 < 0.1
$\hat{\beta}_{public}$ Pr(>|t|) ≈ 0.4272 > 0.1
$\hat{\beta}_{expend}$ Pr(>|t|) ≈ 0.0123 ≈ 0.1

$\hat{\sigma}$ = 26.34, df = *n*-*p*-1 = 43
R^2 ≈ 0.879 ⇒ 87.9% of variability explained

Georgia Tech

4

Testing for Subsets of Coefficients

```
## Compare models: reduced with controlling variables only vs. full with all variables
anova(regression.line)
Analysis of Variance Table

Response: sat
          Df Sum Sq Mean Sq  F value    Pr(>F)
takers     1 181024  181024 260.8380 < 2.2e-16 ***
rank       1  11209   11209  16.1512 0.0002313 ***
income     1   2858    2858   4.1182 0.0486431 *
years      1  16080   16080  23.1701  1.86e-05 ***
public     1    252     252   0.3631 0.5499447
expend     1   4745    4745   6.8369 0.0122629 *
Residuals 43  29842     694

## compute partial-F statistic
fstat = ((2858+16080+252+4745)/4)/(29842/43)
pvalue = 1-pf(fstat,4,43)
pvalue
[1] 3.349778e-05
```

5

Testing for Subsets of Coefficients

Test: $\mathrm{H}_0\text{: } \beta_{income} = \beta_{public} = \beta_{years} = \beta_{expend} = 0$

How were the F-statistic and the p-value computed?

$$\text{F-statistic} = \frac{\mathrm{SS}_{\mathrm{Reg}}(income, public, years, expend \mid takers, rank)/4}{\mathrm{SSE}/(50-6-1)}$$

$$\Pr\left(\mathrm{F}_{4,43} > \text{F-statistic}\right) = 1 - \Pr\left(\mathrm{F}_{4,43} < \text{F-statistic}\right)$$

Interpretation: The p-value is approximately 0, thus reject the null hypothesis. We conclude that at least one other predictor among the four predictors (*income*, *years*, *public* and *expend*) will be significantly associated to the state-average SAT score.

6

Using Residuals to Create Better Rankings

Bias Selection: Some state universities require the SAT and some require a competing exam. States with a high proportion of takers probably have "in state" requirements for the SAT. In states without this requirement, only the more elite students will take the SAT, causing a bias.

```
## Consider model with the two controlling factors to correct for bias
reduced.line = lm(sat ~ takers + rank)
## obtain the order of states by the residuals of the reduced model
order.vec = order(reduced.line$res, decreasing = TRUE)
## Reorder states. Create table including state name, new and old order.
states = factor(data[order.vec, 1])
newtable = data.frame(State = states, Residual = as.numeric(round(reduced.line$res[order.vec],
1)), oldrank = (1:50)[order.vec])
newtable
```

7

Using Residuals to Create Better Rankings

	State	Residual	oldrank
1	Connecticut	53.9	35
2	Iowa	53.5	1
3	New Hampshire	45.8	28
4	Massachusetts	41.9	41
5	New York	40.9	36
6	Minnesota	40.6	7
7	Kansas	35.8	4
8	SouthDakota	33.4	2
⋮			
43	Arkansas	-31.2	12
44	WestVirginia	-38.9	25
45	Nevada	-45.4	30
46	Mississippi	-49.3	16
47	Texas	-50.3	45
48	Georgia	-63.0	49
49	NorthCarolina	-71.3	48
50	SouthCarolina	-98.5	50

After controlling for selection bias, Connecticut moved from 35^{th} to 1^{st}.

After controlling for selection bias, Mississippi moved from 16^{th} to 46^{th}.

8

Summary

9