Topics

- Interpretation of Output
- Multicollinearity
- Variable Selection
Air Pollution Data

- hh/datasets/usair.dat
- Response SO2 measurements in 41 metropolitan areas
- Predictors
  - temp
  - firms
  - popn
  - wind
  - precip
  - rain

Model for log(SO2) as a function of temp, log(firms), log(popn), wind, precip, rain.
Scatterplot Matrix
## Correlations

<table>
<thead>
<tr>
<th></th>
<th>lS02</th>
<th>temp</th>
<th>lfirms</th>
<th>lpopn</th>
<th>wind</th>
<th>precip</th>
<th>rain</th>
</tr>
</thead>
<tbody>
<tr>
<td>lS02</td>
<td>1.00</td>
<td>-0.55</td>
<td>0.34</td>
<td>0.10</td>
<td>-0.01</td>
<td>0.05</td>
<td>0.48</td>
</tr>
<tr>
<td>temp</td>
<td>-0.55</td>
<td>1.00</td>
<td>-0.15</td>
<td>0.05</td>
<td>-0.35</td>
<td>0.39</td>
<td>-0.43</td>
</tr>
<tr>
<td>lfirms</td>
<td>0.34</td>
<td>-0.15</td>
<td>1.00</td>
<td>0.87</td>
<td>0.31</td>
<td>0.03</td>
<td>0.16</td>
</tr>
<tr>
<td>lpopn</td>
<td>0.10</td>
<td>0.05</td>
<td>0.87</td>
<td>1.00</td>
<td>0.27</td>
<td>-0.04</td>
<td>-0.01</td>
</tr>
<tr>
<td>wind</td>
<td>-0.01</td>
<td>-0.35</td>
<td>0.31</td>
<td>0.27</td>
<td>1.00</td>
<td>-0.01</td>
<td>0.16</td>
</tr>
<tr>
<td>precip</td>
<td>0.05</td>
<td>0.39</td>
<td>0.03</td>
<td>-0.04</td>
<td>-0.01</td>
<td>1.00</td>
<td>0.50</td>
</tr>
<tr>
<td>rain</td>
<td>0.48</td>
<td>-0.43</td>
<td>0.16</td>
<td>-0.01</td>
<td>0.16</td>
<td>0.50</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Multiple Regression
HH Chapter 9
Air Pollution Example
Scatter Plot
Added Variable Plots
Interpretation
Multicollinearity
### Multiple Regression

**HH Chapter 9**

**Air Pollution Example**

**Interpretation**
- Residual Standard Error
- $F$ statistic
- $R^2$ and adjusted $R$-squared
- Coefficients and their standard errors (in original units)
- $t$-statistics & $p$-values

---

**Multicollinearity**

```r
summary(poll.lm3) (abbreviated)
```

| Coefficients | Estimate   | Std. Error | t value | Pr(>|t|)   |
|--------------|------------|------------|---------|------------|
| (Intercept)  | 6.7142760  | 1.6475086  | 4.075   | 0.000261 **|
| temp         | -0.0649495 | 0.0227711  | -2.852  | 0.007333 **|
| log(firms)   | 0.3698588  | 0.1934076  | 1.912   | 0.064289 . |
| log(popn)    | -0.1771293 | 0.2335520  | -0.758  | 0.453428   |
| wind         | -0.1738606 | 0.0656713  | -2.647  | 0.012204 * |
| precip       | 0.0156032  | 0.0132718  | 1.176   | 0.247893   |
| rain         | 0.0009153  | 0.0057335  | 0.160   | 0.874104   |

---

Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1

**Residual standard error:** 0.5108 on 34 degrees of freedom

**Multiple R-Squared:** 0.5503,  **Adjusted R-squared:** 0.47

**F-statistic:** 6.936 on 6 and 34 DF,  **p-value:** 7.12e-05
Residual Standard Error

Residual standard error: 0.5108 on 34 degrees of freedom

- Residuals: $e_i = Y_i - \hat{Y}_i$
- Unbiased estimate of $\sigma^2$: $\hat{\sigma}^2 = \sum e_i^2 / (n - p - 1)$
- Sum of squared residuals over degrees of freedom (n - # estimated coefficients in regression)
- Residual standard error is $\hat{\sigma}$
F-statistic: 6.936 on 6 and 34 DF, p-value: 7.12e-05

- Null Hypothesis: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$
- Alternative Hypothesis: at least one coefficient is not zero
- Test statistic:
  \[ F = \frac{\sum (\hat{Y}_i - \bar{Y})^2 / p}{\hat{\sigma}^2} \]
- Distribution: F with \( p \) and \( n - p - 1 \) degrees of freedom under the null hypothesis
R2

Coefficient of Determination or Multiple R-Squared: 0.5503

\[ R^2 = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2} \]

\[ = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}} \]

\[ = \text{Fraction of variation explained by the regression} \]

55% of the variation in log(SO2) can be explained by the linear model
Adjusted R-squared: 0.471

- R2 increases as we add more and more predictors
- R2 = 1 with \( p = n \) (even if the predictors are unrelated to \( Y \! \))
- R2 is only useful to compare models with the SAME response and the SAME number of predictors
- To take into account model complexity

\[
\text{Adjusted } R^2 = 1 - \frac{\hat{\sigma}^2}{S^2_y}
\]
Coefficients

For a 1 unit increase in $X_j$, expect $Y$ to increase by $\hat{\beta}_j$ (with everything else held constant)

$$(1 - \alpha)100\%CI: \hat{\beta}_j \pm t_{\alpha/2, n-p-1}SE(\hat{\beta}_j)$$

Need to be able to interpret after transformation back to original units of SO2

$$\hat{SO2} = e^{\hat{\beta}_0} e^{\hat{\beta}_1 \text{Temp}} e^{\hat{\beta}_2 \log \text{firm}} e^{\hat{\beta}_3 \log \text{popn}} e^{\hat{\beta}_4 \text{wind}} e^{\hat{\beta}_5 \text{precip}} e^{\hat{\beta}_6 \text{rain}}$$

$$\hat{SO2} = e^{\hat{\beta}_0} e^{\hat{\beta}_1 \text{Temp}} \text{firm}^{\hat{\beta}_2} \text{popn}^{\hat{\beta}_3} e^{\hat{\beta}_4 \text{wind}} e^{\hat{\beta}_5 \text{precip}} e^{\hat{\beta}_6 \text{rain}}$$
Temperature

- a 1 unit increase in temperature: \( \exp(-0.065 \times 1) = 0.94 \)
- 95% interval (0.89, 0.98) (\( \exp(\text{CI}) \))
Number of Firms

- a 2 fold increase (doubling) of number of firms: 
  \[ 2^{0.37} = 1.29 \]
- 95% interval (0.98, 1.70) (2CI)
Null hypothesis: $\beta_j = 0$

Alternative hypothesis $\beta_j \neq 0$

test statistic $\hat{\beta}_j / SE(\hat{\beta}_j)$

Distribution: under null, Student-t with $n - p - 1$ degrees of freedom

Conclusion: reject null if p-value $< \alpha$

Does not allow simultaneous tests of two or more coefficients – only one coefficient at a time!
Signs of Multicollinearity

- coefficient of popn in multiple regressions has a negative sign
- coefficient of popn in simple linear regression (correlation) has a positive sign!
- Coefficient is adjusted for other variables so sign change may be meaningful
- Maybe a symptom of multicollinearity
Variance Inflation Factor

\[
\text{vif}(\log(\text{SO2}) \sim \text{temp} + \log(\text{firms}) + \log(\text{popn}) + \text{wind} + \text{precip} + \text{rain},
\]
\[
\text{data=pollution})
\]

\[
\begin{align*}
\text{temp} & \quad l\text{firm} & \quad l\text{popn} & \quad \text{wind} & \quad \text{precip} & \quad \text{rain} \\
4.15 & \quad 5.32 & \quad 5.41 & \quad 1.35 & \quad 3.74 & \quad 3.54
\end{align*}
\]

\[
\text{VIF}(X_j) = 1/(1 - R^2_j) \quad (R^2_j \text{ is percent of variation in } X_j \text{ that can be explained by the other } X's)
\]

Values greater than 5 indicate multicollinearity (redundancy and instability)

Variance of coefficient is inflated by VIF over variances with no multicollinearity present (SE is inflated by square root of VIF)
Problems with Multicollinearity

- Variables may appear to be unimportant (when they are) (large VIF may lead to larger SE’s and smaller \( t \) statistics)
- Coefficient estimates are unstable and hard to interpret (can estimate combinations of coefficients but not individual coefficients)
Dealing with Multicollinearity

- Collect more data to reduce correlation among explanatory variables
- Keep all variables and drop interpretation of coefficients (focus on prediction)
- Eliminate redundant variables (subjective)
- Automatic Variable selection
- Ridge Regression or other Shrinkage estimates