````
---
title: 'Multiple Regression: variable selection'
output:
  html_document: default
  word_document: default
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

<font size = "6">
</br></br>


- In a general setting with response $Y$ and $p$ predictors $x_{1},\dots,x_{p}$ we will
typically begin a regression analysis with two primary steps: </br></br>
1. Exploring the suitability of a linear relationship between the response and each
predictor.</br></br>
2. Selecting which variables should be included in the regression model.

</br></br></br></br>

- The problem of selecting variables to include in a regression model is a fundamental and
important statistical problem. The set of possible regression models has $2^{p}$ elements.

</br></br></br></br>

- Many methods have been proposed for variable selection in regression. We will consider
one simple and practical approach known as backward selection.

</br></br></br></br>

- The general strategy that we will adopt is to start with a full model containing all main
effects and some low order interactions and sequentially remove terms that are not
significant one-by-one, sequentially re-fitting the model until all remaining terms are
significant.

</br></br></br></br>

- We will prefer to remove higher order interactions whenever possible as this simplifies
the model interpretation.

</br></br></br></br>

- As a rough rule of thumb, if the sample size is $n$ we should ensure that no more than
$n/3$ terms are included in any model considered to ensure that the number of parameters is
not excessive. Alternatives to this restriction are possible.

</br></br></br></br>

- Example: Air Pollution Study
</br></br>

- The data comprise $n = 110$ observations of ozone concentration and the objective is to
relate ozone concentration to predictors: wind speed, air temperature, intensity of solar
radiation.
</br></br></br></br>

```{r}
ozone.pollution<-read.table(file
='~/Desktop/stat359/data/ozone.data.txt',header=TRUE,sep="")
attach(ozone.pollution)
names(ozone.pollution)
````
````

````
library(knitr)
kable(ozone.pollution, caption = 'Ozone Study',align='l')
pairs(ozone.pollution)
```

</br></br></br></br>

- It looks like ozone may have a positive relationship with intensity of solar radiation
and a positive relationship with temperature, while it appears to have a negative
relationship with wind speed.

</br></br></br></br>

- There may be some curvature in the relationship between ozone and air temperature as well
as with wind speed. A quadratic function may be sufficient in both cases.

</br></br></br></br>

- Start with a model having quadratic terms for all three factors; include all of the three
2-way interactions $rad \times temp$, $rad \times wind$, $temp \times wind$; and the single
3-way interaction $rad \times temp \times wind$.

</br></br></br></br>

- In total the model we start with will have 3 main effects, 3 quadratic terms, 3 2-way
interactions, 1 3-way interaction and an intercept leading to 11 parameters. This is fewer
than 110/3 parameters so there is no concern about over-parameterization.

</br></br></br></br>

```{r}
model1<-lm(ozone~temp*wind*rad+I(rad^2)+I(temp^2)+I(wind^2))
summary(model1)
```

</br></br></br></br>

- The initial model has an $R^{2} = 0.739$ (this is the highest it will be) and the 3-way
interaction term is not significant (p-value = 0.514).

</br></br></br></br>

```{r}
model2<-update(model1,.~.- temp:wind:rad)
summary(model2)
```

</br></br>

- $temp \times rad$ is the least significant 2-way interaction

</br></br>

```{r}
model3<-update(model2,.~.- temp:rad)
summary(model3)
```

</br></br>

- $temp \times wind$ is borderline and is removed

</br></br>
````
````

```{r}
model4<-update(model3,.~.- temp:wind)
summary(model4)
```

</br></br>

- $wind \times rad$ is not significant at the 0.05 level and we remove it

</br></br>

```{r}
model5<-update(model4,.~.- wind:rad)
summary(model5)
```

</br></br>

- Remove the quadratic term for solar radiation (p-value = 0.422).

</br></br>

```{r}
model6<-update(model5,.~.- I(rad^2))
summary(model6)
```

</br></br>

- All remaining terms are significant. We examine some model diagnostics.

</br></br>

```{r}
par(mfrow=c(1,3))
plot(model6,which=c(1,2,4))
```

</br></br>

- It appears as though the variance of the residuals is not constant and increases with fitted values. The distribution of the residuals also may be right skewed (long right tail).

</br></br>

- Apply a log transformation to the response.
</br></br>

```{r}
model7<-update(model6,log(.)~.)
summary(model7)
```
</br></br>

- After the log transformation the quadratic term for air temperature can be removed (p-value = 0.493).
</br></br>

```{r}
model8<-update(model7,~.-I(temp^2))
summary(model8)

```

</br></br>

````
```{r}
par(mfrow=c(1,3))
plot(model8,which=c(1,2,4))
```
````

</br></br>

- The variance of the residuals appears constant and their distribution does not appear to deviate from normality. Observation 17 looks as though it may be potentially influential. This observation appears extreme in all of the diagnostic plots.
</br></br>

````
```{r}
## examine the sensitivty of the results
model9<-update(model8,.~.,subset=(1:length(ozone)!=17))
summary(model9)
```
````

</br></br>

- The coefficient estimates (aside from the intercept) appear stable and the R^2 is also stable to removal of observation 17.

</br></br>

- In summary, all three factors, wind speed, air temperature, intensity of solar radiation appear related to ozone concentration.
</br></br>

- The effects are additive and these factors do not appear to interact in their relationship with ozone concentration.
</br></br>

- On the log-scale, air temperature and solar radiation are positively related to ozone concentration while wind speed is quadratically related to wind speed.
</br></br>

- Final model:
$$
\hat{\text{Ozone}} = \exp\{0.72 + 0.046 \times \text{TEMP} + 0.0025 \times \text{RAD} -0.22 \times \text{WIND} + 0.0072 \times \text{WIND}^{2} \}
$$
</br></br>

- This model explains roughly 69% of the variability in ozone concentration in this dataset.