Applied Statistical Regression
HS 2011 – Week 09

Marcel Dettling

Institute for Data Analysis and Process Design

Zurich University of Applied Sciences

marcel.dettling@zhaw.ch

http://stat.ethz.ch/~dettling

ETH Zürich, November 21, 2011

Applied Statistical Regression
HS 2011 – Week 09

Collinearity = Correlated Predictors

If two or more predictors are strongly correlated, i.e. try to explain very similar aspects of the data, estimation is difficult. The regression coefficients will be less precisely estimated, which influences interpretation of the results.

There is a need to recognize collinearity!

1) *Plot the correlation matrix of the predictors*

```
plotcorr(cor(my.dat))
```

2) *Variance Inflation Factors*

$$Var(\hat{\beta}_k) = \sigma_E^2 \cdot \frac{1}{1-R_k^2} \cdot \frac{1}{\sum_{i=1}^{n}(x_{ik} - \overline{x}_k)}$$

Applied Statistical Regression
HS 2011 – Week 09

How to Deal with Correlated Predictors?

1) **Generate new variables**

→ **see example on next slides…**

2) **Variable selection**

Only work with the relevant variables, and omit the redundant ones. This often helps a lot. We will be discussing variable selection in detail.

3) **The Lasso and Ridge Regression**

These are penalized regression methods, which sparsely spend degrees of freedom. To be discussed later.

Applied Statistical Regression
HS 2011 – Week 09

Example

Understanding how car drivers adjust their seat would greatly help engineers to design better cars. Thus, the measured

hipcenter = horizontal distance of hips to steering wheel

and tried to explain it with several predictors, namely:

Age	age in years
Weight	weight in pounds
HtShoes, Ht, Seated	height w/o, w/ shoes, seated height
Arm, Thigh, Leg	arm, thigh and leg length

We first fit a model with all these (correlated!) predictors

Applied Statistical Regression
HS 2011 – Week 09

Example: Fit with All Predictors

```
> library(faraway); data(seatpos)
> summary(lm(hipcenter~., data=seatpos))

             Estimate Std. Error t value Pr(>|t|)
(Intercept) 436.43213  166.57162   2.620   0.0138 *
Age           0.77572    0.57033   1.360   0.1843
Weight        0.02631    0.33097   0.080   0.9372
HtShoes      -2.69241    9.75304  -0.276   0.7845
Ht            0.60134   10.12987   0.059   0.9531
Seated        0.53375    3.76189   0.142   0.8882
Arm          -1.32807    3.90020  -0.341   0.7359
Thigh        -1.14312    2.66002  -0.430   0.6706
Leg          -6.43905    4.71386  -1.366   0.1824

Residual standard error: 37.72 on 29 degrees of freedom
Multiple R-squared: 0.6866,  Adjusted R-squared: 0.6001
F-statistic:  7.94 on 8 and 29 DF,  p-value: 1.306e-05
```

Example: Generating New Variables

The body height is certainly a key predictors when it comes to the position of the driver seat. We leave this as it was, and change several of the other predictors:

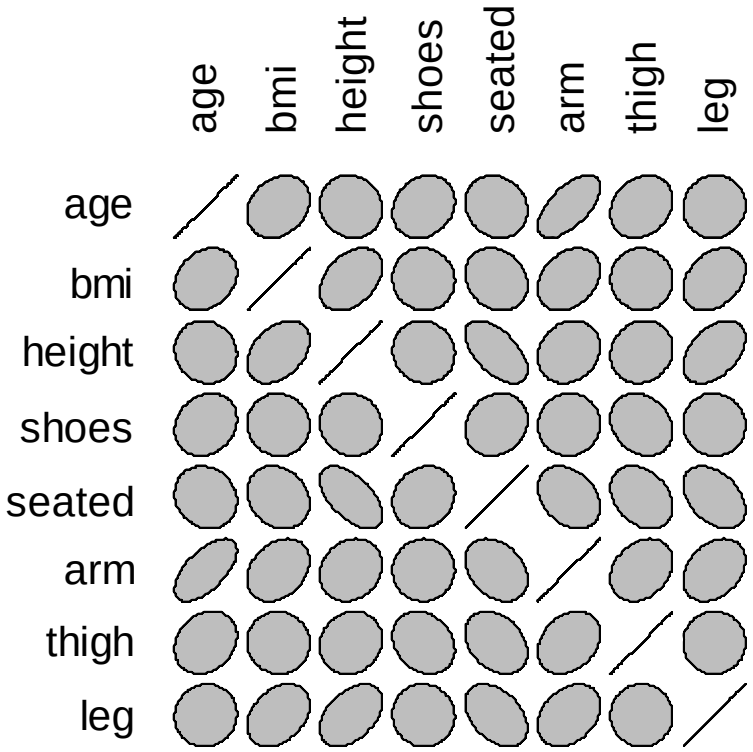
```
age    <- Age
bmi    <- (Weight*0.454)/(Ht/100)^2
shoes  <- HtShoes-Ht
seated <- Seated/Ht
arm    <- Arm/Ht
thigh  <- Thigh/Ht
leg    <- Leg/Ht
```

Does this solve the correlation problem...?

Applied Statistical Regression
HS 2011 – Week 09

Example: New Correlation Matrix

We visualize again using function `plotcorr()`



Applied Statistical Regression
HS 2011 – Week 09

Example: Fit with New Predictors

```
> summary(lm(hipc~., data=new.seatpos))

             Estimate Std. Error t value Pr(>|t|)
(Intercept) -632.0063   490.0451  -1.290    0.207
age           -0.7402     0.5697  -1.299    0.204
bmi           -0.4234     2.2622  -0.187    0.853
height         3.6521     0.7785   4.691 5.98e-05 ***
shoes          2.6964     9.8030   0.275    0.785
seated      -171.9495   631.3719  -0.272    0.787
arm          180.7123   655.9536   0.275    0.785
thigh        141.2007   443.8337   0.318    0.753
leg         1090.0111   806.1577   1.352    0.187

Residual standard error: 37.71 on 29 degrees of freedom
Multiple R-squared: 0.6867, Adjusted R-squared: 0.6002
F-statistic: 7.944 on 8 and 29 DF,  p-value: 1.3e-05
```

Applied Statistical Regression
HS 2011 – Week 09

Variable Selection: Why?

We want to fit a regression model…

Case 1: functional form and predictors exactly known
→ *estimation, test, confidence and prediction intervals*

Case 2: neither functional form nor the predictors are known
→ *explorative model search among potential predictors*

Case 3: we are interested in only 1 predictor, but want to correct for the effect of other covariates
→ *which covariates we need to correct for?*

Question in cases 2 & 3: WHICH PREDICTORS TO USE?

Applied Statistical Regression
HS 2011 – Week 09

Variable Selection: Technical Aspects

We want to keep a model small, because of

1) **Simplicity**
   → *among several explanations, the simplest is the best*

2) **Noise Reduction**
   → *unnecessary predictors leads to less accuracy*

3) **Collinearity**
   → *removing excess predictors facilitates interpretation*

4) **Prediction**
   → *less variables, less effort for data collection*

Method or Process?

- Please note that variable selection is not a method. The search for the best predictor set is an iterative process, which also involves estimation, inference and model diagnostics.
- For example, outliers and influential data points will not only change a particular model – they can even have an impact on the model we select. Also variable transformations will have an impact on the model that is selected.
- Some iteration and experimentation is often necessary for variable selection. The ultimate aim is finding a model that is smaller, but as good or even better than the original one.

Applied Statistical Regression
HS 2011 – Week 09

Example: Variable Selection

```
> summary(lm(hipc~., data=new.seatpos))

             Estimate Std. Error t value Pr(>|t|)
(Intercept) -632.0063   490.0451  -1.290    0.207
age           -0.7402     0.5697  -1.299    0.204
bmi           -0.4234     2.2622  -0.187    0.853
height         3.6521     0.7785   4.691 5.98e-05 ***
shoes          2.6964     9.8030   0.275    0.785
seated      -171.9495   631.3719  -0.272    0.787
arm          180.7123   655.9536   0.275    0.785
thigh        141.2007   443.8337   0.318    0.753
leg         1090.0111   806.1577   1.352    0.187

Residual standard error: 37.71 on 29 degrees of freedom
Multiple R-squared: 0.6867, Adjusted R-squared: 0.6002
F-statistic: 7.944 on 8 and 29 DF,  p-value: 1.3e-05
```

Applied Statistical Regression
HS 2011 – Week 09

Backward Elimination

→ **Removing more than one variable at a time is problematic**

- Start with the full model, and exclude the predictor with the highest p-value, as long as it exceeds α_{crit}
- Refit the model with the reduced predictor set. Again exclude the least significant predictor if it exceeds α_{crit}
- Repeat until all “non-significant” predictors are removed. Then, we stop and have found the final model.

→ Usually $\alpha_{crit} = 0.05$, for prediction also 0.15 or 0.20

→ **R-demo…**

Applied Statistical Regression
HS 2011 – Week 09

Backward Elimination: Final Result

```
> summary(lm(hipc ~ age + height + leg, data=new.seatpos)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -696.6107   136.5691  -5.101 1.27e-05 ***
age           -0.5835     0.3789  -1.540    0.133
height         3.7712     0.5800   6.502 1.94e-07 ***
leg         1141.0702   690.5917   1.652    0.108

Residual standard error: 35.1 on 34 degrees of freedom
Multiple R-squared: 0.6817,	Adjusted R-squared: 0.6536
F-statistic: 24.27 on 3 and 34 DF,  p-value: 1.403e-08
```

Interpretation of the Result

- The remaining predictors are now “more significant” than before. This is almost always the case. Do not overestimate the importance of these predictors!
- Collinearity among the predictors is usually at the root of this observation. The predictive power is first spread out among several predictors, then it becomes concentrated.
- **Important**: the removed variables can still be related to the response. If we run a simple linear regression, they can even be significant. In the multiple linear model however, there are other, better, more informative predictors.

Applied Statistical Regression
HS 2011 – Week 09

Forward Selection

→ This is an analogue to the backward elimination.

- Starts with an empty model, i.e. a model where only the intercept, but no predictors are present.
- We try all predictors one after the other and add the one which has the lowest p-value, provided it's below α_{crit}.
- Repeat until either all predictors are included in the model, or no further significant predictors can be found.

→ Feasible in situation where $p > n$

→ Computationally cheap, thus historically popular

Applied Statistical Regression
HS 2011 – Week 09

Stepwise Regression

→ This is mix between forward and backward selection.

- Forward and backward steps are carried out alternately. One can either start with the full model (1st step backward), or with the empty model (1st step forward).
- In each forward step all predictors can be included in the model, even those that were removed in a previous step.
  In the backward steps, any predictor can be removed.
- Decisions can be based on p-values of the hypothesis tests.

→ Often applied, is the default in R function `step()`

→ However, `step()` is based on AIC/BIC, not on p-values

Applied Statistical Regression
HS 2011 – Week 09

Testing Based Variable Selection

What are the drawbacks of the forward, backward or stepwise approach if based on the p-values of hypothesis tests?

1) **Missing the „best“ model**
   → *due to „one-at-a-time“ adding/dropping*

2) **Multiple testing problem**
   → *p-values should not be taken too literally*

3) **Missing link to final objective**
   → *hypothesis tests != prediction/explanation*

4) **Too small models**
   → *for prediction, bigger models usually perform better*

Applied Statistical Regression
HS 2011 – Week 09

All Subsets Regression

If there are *m* potential predictors, we can build 2^m models.

- a complete, exhaustive search over all these models is naturally only feasible if *m* is reasonably small

We need a means of comparison for models of different size!

1) Coefficient of determination R^2
2) Test statistic or p-value of the global F-test
3) Estimated error variance $\hat{\sigma}_E^2$

→ *all these are measuring goodness-of-fit: they improve when more predictors are added to the model!*

Applied Statistical Regression
HS 2011 – Week 09

AIC/BIC

Bigger models are not necessarily better than smaller ones!

→ *balance goodness-of-fit with the number of predictors used*

AIC Criterion:

$$\begin{aligned} AIC &= -2\max(\log likelihood) + 2p \\ &= const + n\log(RSS/n) + 2p \end{aligned}$$

BIC Criterion:

$$\begin{aligned} BIC &= -2\max(\log likelihood) + p\log n \\ &= const + n\log(RSS/n) + p\log n \end{aligned}$$

AIC or BIC?

Both can lead to similar decisions, but BIC punishes larger models more heavily:

→ **BIC models tend to be smaller!**

- AIC/BIC is not limited to all subset regression
- Criteria can also be (and are!) applied in the backward, forward or stepwise approaches.
- In R, variable selection is generally done by function `step()`
- Default choice: stepwise regression with AIC as a criterion.

Final Remark

- Every procedure may yield a different “best” model.
- If we could obtain another sample from the same population, even a fixed procedure might result in another “best” model.
- “Best model”: element of chance, “random variable”

How can we mitigate this in practice?

It is usually advisable to not only consider the “best” model according to a particular procedure, but to check a few more models that did nearly as good, if they exist.

Applied Statistical Regression
HS 2011 – Week 09

Model Selection with Hierarchical Input

→ Some regression models have a natural hierarchy.

I.e. in polynomial models, x^2 is a higher order term than x

Important:

Lower order terms should not be removed from the model before higher order terms in the same variable. As an example, consider the polynomial model:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

→ **see blackboard…**

Applied Statistical Regression
HS 2011 – Week 09

Interactions and Categorical Input

Models with Interactions

Do not remove main effect terms if there are interactions with these predictors contained in the model.

Categorical Input

- If a single dummy coefficient is non-significant, we cannot just kick this term out of the model, but we have to test the entire block of indicator variables.
- When we work manually and testing based, this will be done with a partial F-test. When working criterion based, `step()` does the right thing