

# Homework 0

My Name

## 1 Model Building

### 1.1 Data Preprocessing

**Multicollinearity Analysis:** Generate a correlation diagram of the covariates to investigate the relationship between variables. As shown in Figure 1a and Table 1, ...

**Transformation and Residual Analysis:** The next step is to check model assumptions and transformation.

### 1.2 Model Summary

The estimated model is shown as below. The residual standard error is 0.3 and the AIC value is -67. The detail of the summary is shown in Output 1.

$$\begin{aligned}\tilde{y} &= 12 + 0.2 * V5 \\ &\quad - 0.01 * V15 : V12 \\ y &= (75\tilde{y} + 1)^{2/3} - 20\end{aligned}$$

## 2 Important Variables and Reliability Assessment

Combing the result of model comparison, the final model ...

## 3 Model Comparison

Compare the proposed model **m1** with candidate models: regression tree, random forest, bagging [Faraway(2016)], ... The optimal lo and gam model are as follows:

```
> mod.lo$call
loess(formula = y ~ V11 + V12 , data = dat2, subset = ss,
      span = 0.5, degree = 2)

> mod.gam$call
gam(formula = y ~ s(V5) + s(V12) + s(V17), data = dat2,
    subset = ss)
```

### Problem 1

(a).

```
> # library(oehlert)
> library(faraway)
> library(MASS)
> # m<-lm(durability ~ brand, data = ex11.3)
> # boxcox(m)
> # anova(m)
```

```
> n<-6;g<-5  
> l<-(80.45*qt(0.005,25,4)-1)/n;u<-(80.45*qt(0.995,25,4)-1)/n  
> l;u
```

```
[1] 2.606462
```

```
[1] 268.0322
```

I don't believe this interval has 99% coverage because the assumption may not be satisfied. Brand 1 is very different from other brands, which means  $\alpha_1$  may not follow a normal distribution.

(b).

```
[1] 0.7930562
```

```
[1] 0.9912905
```

## References

[Faraway(2016)] J. J. Faraway, *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC, 2016.

# Appendices

## Appendix A: Figures

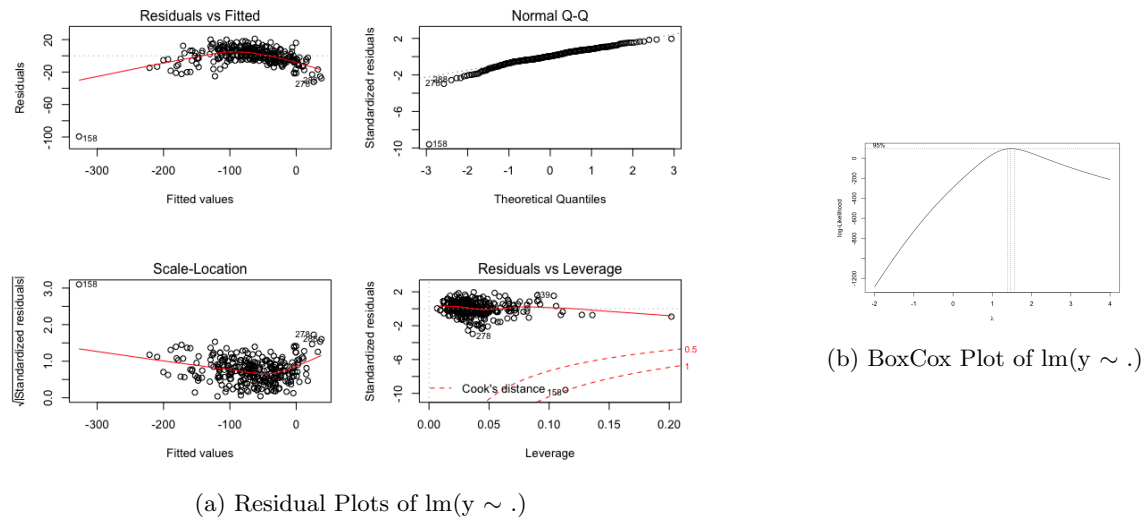


Figure 1: Diagnostics Plots of  $\text{lm}(y \sim .)$

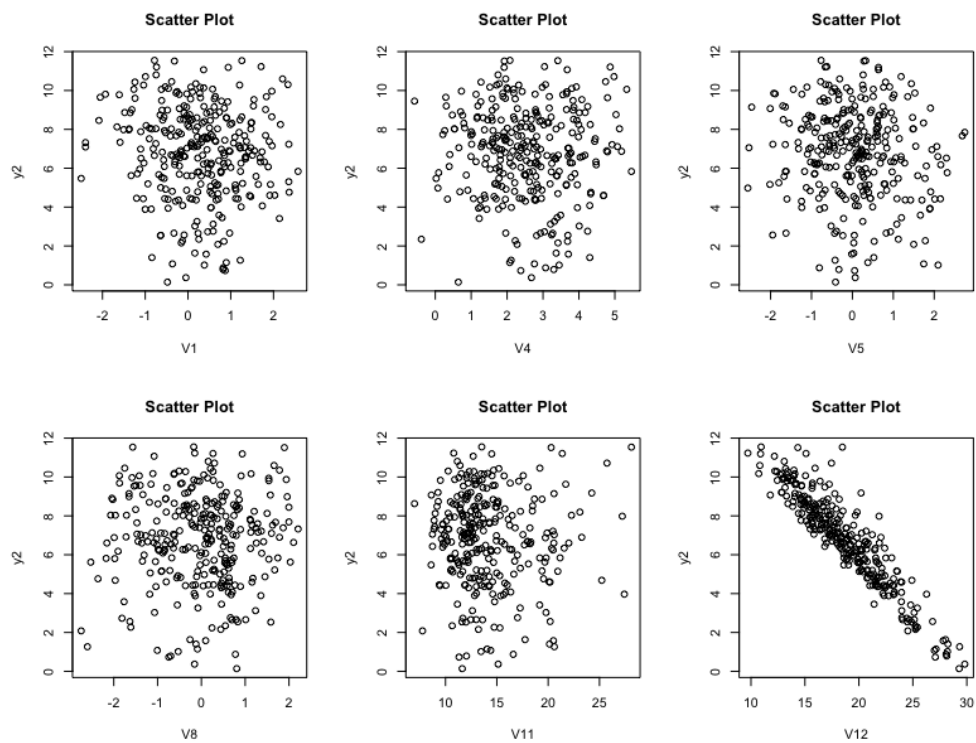


Figure 2: Scatter plot between response and variables

## Appendix B: Tables

Table 1: Grouping of Variables

$G0$	1	2	3	4
$G1$	5	6	7	

Table 2: Variables selected and VSD values

<i>method</i>	<i>Variables</i>	<i>ARM</i>			<i>BIC</i>		
		<i>VSD</i>	<i>VSD_minus</i>	<i>VSD_plus</i>	<i>VSD</i>	<i>VSD_minus</i>	<i>VSD_plus</i>
LASSO	{1 2 3 4}	0	0	0	0	0	0
SCAD	{1 2 3}	0	0	0	0	0	0
MCP	{1 2}	0	0	0	0	0	0

Table 3: Variable Importance

Metric	Importance Order
IncMSE	1 2 3 4
IncNodePurity	1 2 3 4
SOIL	1 2 3 4
	(leftmost is the most important)

Table 4: Uncertainty Assessment

<i>Method</i>		S={1, 2, 3, 4}
<i>Instability</i>	Sequential	0
	Bootstrap	0
	Perturbation	0
<i>ARM</i>	<i>VSD</i>	0
	<i>VSD_minus</i>	0
	<i>VSD_plus</i>	0
	<i>F-measure</i>	0
	<i>G-measure</i>	0
<i>BIC</i>	<i>VSD</i>	0
	<i>VSD_minus</i>	0
	<i>VSD_plus</i>	0
	<i>F-measure</i>	0
	<i>G-measure</i>	0

Table 5: Model Comparison

Comparison	Winner	Winning Fraction of m1
m1 vs. Regression Tree	m1	1
m1 vs. random forest	m1	1
m1 vs. bagging	m1	1
m1 vs. loess	m1	1
m1 vs. gam	gam	0

Table 6: Cross Validation MSE and Absolute Error

Model	CV MSE	CV Mean Absolute Error
m1	0	0
Regression Tree	0	0
random forest	0	0
bagging	0	0
loess	0	0
gam	0	0

## Appendix C: R output

### Output 1

```
> summary(m3.1)
```

Coefficients:

```

      Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.915676   0.502463  25.705 < 2e-16 ***
V5           0.287077   0.108717   2.641 0.00873 **
V8          -0.033139   0.020466  -1.619 0.10649
V11          0.402557   0.042078   9.567 < 2e-16 ***
---

```

Residual standard error: 0.3505 on 289 degrees of freedom

```
> extractAIC(m3.1)
```

```
[1] 10.0000 -617.0269
```