Relaxing the Model Assumptions

Standardizing residuals

We recall from previous results that for residual random vectors $\mathbf{Y} - \widehat{\mathbf{Y}}$ we have (under correct specification)

$$\mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y} - \widehat{\mathbf{Y}}|\mathbf{X}] = \sigma^2(\mathbf{I}_n - \mathbf{H}).$$

Taking the diagonal elements, this implies that the variance of the ith residual is

$$\sigma^2(1 - h_{ii})$$

which we may use as a means to process the residual so that it appears on a standard scale.

Outliers

An **outlier** is a point for which the residual (or standardized residual) is large.

- Such points need to be considered carefully as they may exert a lot of influence on the fit.
- Outliers may need to be deleted from the data set.

Using standard large sample arguments, a data point may be considered an outlier if

$$\left| \frac{y_i - \widehat{y}_i}{\sqrt{\sigma^2(1 - h_{ii})}} \right| > 2$$

Deletion residuals

Consider the fit of a regression model to data indexed $i = 1, \ldots, n$, and consider refitting the model with the ith point deleted. Let

- $\mathbf{y}_{(i)}$ be the response vector with the ith response deleted;
- $\mathbf{X}_{(i)}$ be the $\mathbf{X}$ matrix with the ith row deleted.

The least squares estimate when point i is deleted is

$$\widehat{\beta}_{(i)} = (\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)})^{-1} \mathbf{X}_{(i)}^\top \mathbf{y}_{(i)}.$$

We then have the prediction at $\mathbf{x} = \mathbf{x}_i$ as

$$\widehat{y}_{(i)} = \mathbf{x}_i \widehat{\beta}_{(i)}.$$

We attempt to assess model validity using this 'out-of-sample' prediction.

Deletion residuals (cont.)

The ith PRESS (**Pre**diction **S**um of **S**quares) residual is defined as

$$e_{(i)} = y_i - \widehat{y}_{(i)} = y_i - \mathbf{x}_i \widehat{\beta}_{(i)} = y_i - \mathbf{x}_i (\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)})^{-1} \mathbf{X}_{(i)}^\top \mathbf{y}_{(i)}$$

We compute this quantity for $i = 1, \ldots, n$. It transpires (see Appendix) that

$$e_{(i)} = \frac{e_i}{1 - h_{ii}}$$

where h_{ii} is the ith diagonal element of hat matrix $\mathbf{H}$.

Deletion, Leverage and influence

We can extend the idea of deletion for residuals to deletion for inference: we compare estimates

- $\widehat{\beta}$ from the full data set
- $\widehat{\beta}_{(i)}$ when the ith data point is removed.

As well as the regression estimates, we also have the estimates of σ^2:

- $\widehat{\sigma}^2$ from the full data set
- $\widehat{\sigma}^2_{(i)}$ when the ith data point is removed.

Deletion, Leverage and influence (cont.)

We might use for data point i

$$D_i = \frac{(\widehat{\beta}_{(i)} - \widehat{\beta})^\top (\mathbf{X}^\top \mathbf{X})(\widehat{\beta}_{(i)} - \widehat{\beta})}{p\mathrm{MS}_{\mathrm{Res}}} = \frac{(\widehat{\mathbf{y}}_{(i)} - \widehat{\mathbf{y}})^\top (\widehat{\mathbf{y}}_{(i)} - \widehat{\mathbf{y}})}{p\mathrm{MS}_{\mathrm{Res}}}$$

as a global measure of influence on inference on a standardized scale.

D_i is **Cook's distance.**

Leverage

From standard theory, we have that

$$\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

where $\mathbf{H}$ is the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$. Thus

$$\widehat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$$

The coefficients h_{ij} measure the importance of each of the original data $y_1, \ldots, y_n$ in predicting y_i.

h_{ij} is termed the **leverage** of point j on point i. We have

$$h_{ii} = \mathbf{x}_i (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i^\top$$

and if this value is large, the data point i is considered influential.

Example: Life Cycle Data

From the `R` help:

Under the life-cycle savings hypothesis as developed by Franco Modigliani, the savings ratio is explained by per-capita disposable income, the percentage rate of change in per-capita disposable income, and two demographic variables: the percentage of population less than 15 years old and the percentage of the population over 75 years old. The data are averaged over the decade 1960–1970 to remove the business cycle or other short-term fluctuations.

- predictor `pop15` – % of population under 15
- predictor `pop75` – % of population over 75
- predictor `dpi` – real per-capita disposable income
- predictor `ddpi` – % growth rate of dpi
- response `sr` – savings ratio (aggregate personal saving divided by disposable income)

Analysis

```
> LifeCycleSavings
                    sr pop15 pop75     dpi  ddpi
Australia        11.43 29.35  2.87 2329.68  2.87
Austria          12.07 23.32  4.41 1507.99  3.93
Belgium          13.17 23.80  4.43 2108.47  3.82
Bolivia           5.75 41.89  1.67  189.13  0.22
Brazil           12.88 42.19  0.83  728.47  4.56
Canada            8.79 31.72  2.85 2982.88  2.43
Chile             0.60 39.74  1.34  662.86  2.67
China            11.90 44.75  0.67  289.52  6.51
     .               .     .     .       .     .
     .               .     .     .       .     .
     .               .     .     .       .     .
Tunisia           2.81 46.12  1.21  249.87  1.13
United Kingdom    7.81 23.27  4.46 1813.93  2.01
United States     7.56 29.81  3.43 4001.89  2.45
Venezuela         9.22 46.40  0.90  813.39  0.53
Zambia           18.56 45.25  0.56  138.33  5.14
Jamaica           7.72 41.12  1.73  380.47 10.23
Uruguay           9.24 28.13  2.72  766.54  1.88
Libya             8.89 43.69  2.07  123.58 16.71
Malaysia          4.71 47.20  0.66  242.69  5.08
```

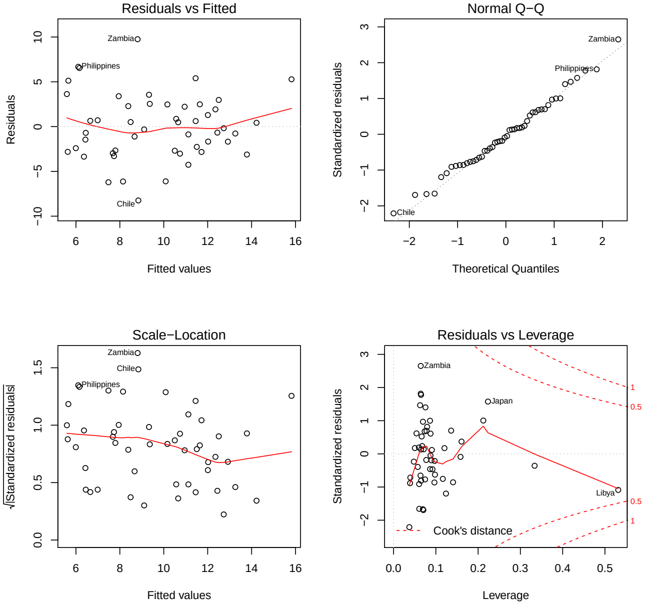
Analysis

```
> str(LifeCycleSavings)
'data.frame':   50 obs. of  5 variables:
 $ sr   : num  11.43 12.07 13.17 5.75 12.88 ...
 $ pop15: num  29.4 23.3 23.8 41.9 42.2 ...
 $ pop75: num  2.87 4.41 4.43 1.67 0.83 2.85 1.34 0.67 1.06 1.14 ...
 $ dpi  : num  2330 1508 2108 189 728 ...
 $ ddpi : num  2.87 3.93 3.82 0.22 4.56 2.43 2.67 6.51 3.08 2.8 ...

> fit1 <- lm(sr ~ pop15 + pop75 + dpi + ddpi,
 +            data = LifeCycleSavings)
> summary(fit1)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
pop15       -0.4611931  0.1446422  -3.189 0.002603 **
pop75       -1.6914977  1.0835989  -1.561 0.125530
dpi         -0.0003369  0.0009311  -0.362 0.719173
ddpi         0.4096949  0.1961971   2.088 0.042471 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-squared:  0.3385,    Adjusted R-squared:  0.2797
F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

Residual Plots and Diagnostics `plot(fit1)`



Diagonals of the hat matrix (h_{ii})

```
inf.diags<-lm.influence(fit1)
data.frame(hat=inf.diags$hat[c(1:7,42:50)])
                       hat
Australia      0.06771343
Austria        0.12038393
Belgium        0.08748248
Bolivia        0.08947114
Brazil         0.06955944
Canada         0.15840239
Chile          0.03729796
Tunisia        0.07456729
United Kingdom 0.11651375
United States  0.33368800
Venezuela      0.08628365
Zambia         0.06433163
Jamaica        0.14076016
Uruguay        0.09794717
Libya          0.53145676
Malaysia       0.06523300
```

Deletion change in β estimates $\widehat{\beta}_{(i)} - \widehat{\beta}_i$

```
> data.frame(signif(inf.diags$coef[c(1:7,42:50),],4))
                Intercept      pop15    pop75        dpi       ddpi
Australia         0.09158 -0.0015260 -0.02905  4.267e-05 -3.157e-05
Austria          -0.07473  0.0008692  0.04474 -3.456e-05 -1.623e-03
Belgium          -0.47520  0.0075010  0.13170 -3.255e-05 -1.435e-03
Bolivia           0.04295 -0.0018570 -0.02467  2.998e-05  8.061e-03
Brazil            0.66040 -0.0089200 -0.19420  1.118e-04  1.344e-02
Canada            0.04023 -0.0009873  0.01118 -3.324e-05 -5.255e-04
Chile            -1.40000  0.0183200  0.22740 -1.776e-05  2.248e-02
Tunisia           0.54500 -0.0152600 -0.08411  4.152e-05  2.031e-02
United Kingdom    0.34520 -0.0052090 -0.18650  1.175e-04  1.978e-02
United States     0.51320 -0.0106500  0.04098 -2.192e-04 -6.485e-03
Venezuela        -0.37380  0.0145800 -0.03648  1.058e-04 -2.442e-02
Zambia            1.11800 -0.0106400 -0.34120  8.136e-05  4.160e-02
Jamaica           0.80830 -0.0145400 -0.06219 -6.566e-06 -5.814e-02
Uruguay          -0.99250  0.0187600  0.03222  1.231e-04  1.967e-02
Libya             4.04200 -0.0697500 -0.41060 -1.800e-05 -2.006e-01
Malaysia          0.27200 -0.0088760  0.03519 -4.633e-05 -1.424e-02
```

Influence Diagnostics

	dfb.1_	dfb.pp15	dfb.pp75	dfb.dpi	dfb.ddpi	dffit	cov.r	cook.d	hat
Australia	0.012	-0.010	-0.027	0.045	0.000	0.063	1.193	0.001	0.068
Austria	-0.010	0.006	0.041	-0.037	-0.008	0.063	1.268	0.001	0.120
Belgium	-0.064	0.051	0.121	-0.035	-0.007	0.188	1.176	0.007	0.087
Bolivia	0.006	-0.013	-0.023	0.032	0.041	-0.060	1.224	0.001	0.089
Brazil	0.090	-0.062	-0.179	0.120	0.068	0.265	1.082	0.014	0.070
Canada	0.005	-0.007	0.010	-0.035	-0.003	-0.039	1.328	0.000	0.158
Chile	-0.199	0.133	0.220	-0.020	0.120	-0.455	0.655	0.038	0.037
Tunisia	0.074	-0.105	-0.077	0.044	0.103	-0.218	1.131	0.010	0.075
United Kingdom	0.047	-0.036	-0.171	0.126	0.100	-0.272	1.189	0.015	0.117
United States	0.069	-0.073	0.037	-0.233	-0.033	-0.251	1.655	0.013	0.334
Venezuela	-0.051	0.101	-0.034	0.114	-0.124	0.307	1.095	0.019	0.086
Zambia	0.164	-0.079	-0.339	0.094	0.228	0.748	0.512	0.097	0.064
Jamaica	0.110	-0.100	-0.057	-0.007	-0.295	-0.346	1.200	0.024	0.141
Uruguay	-0.134	0.129	0.030	0.131	0.100	-0.205	1.187	0.009	0.098
Libya	0.551	-0.483	-0.380	-0.019	-1.024	-1.160	2.091	0.268	0.531
Malaysia	0.037	-0.061	0.032	-0.050	-0.072	-0.213	1.113	0.009	0.065

Influence Diagnostics

- `rstandard`: Standardized residual

$$\frac{y_i - \widehat{y}_i}{\widehat{\sigma}}$$

- `rstudent` Studentized residual

$$\frac{y_i - \widehat{y}_i}{\widehat{\sigma}\sqrt{1 - h_{ii}}}$$

Let

$$\mathbf{C} = (\mathbf{X}^\top \mathbf{X})^{-1} \qquad \mathbf{R} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$$

where $\mathbf{R}$ has rows $\mathbf{r}_i$, $i = 1, \ldots, n$, and

$$\widehat{\sigma}^2_{(i)} = \frac{1}{n-p-1} \sum_{j \neq i} (y_j - \widehat{y}_{j(i)})^2$$

is computed for the data set with point i omitted.

Influence Diagnostics (cont.)

- `dfbetas`: standardized change in coefficient when ith point is deleted; for $j = 1, \ldots, p$,

$$DFBETAS_{j,i} = \frac{\widehat{\beta}_j - \widehat{\beta}_{j(i)}}{\sqrt{\widehat{\sigma}^2_{(i)} C_{jj}}}.$$

We consider a point i to be worthy of examination if

$$|DFBETAS_{j,i}| > \frac{2}{\sqrt{n}}$$

Influence Diagnostics (cont.)

- `dffits`: standardized change in fitted value when ith point is deleted; for $i = 1, \ldots, n$,

$$DFFITS_i = \frac{\widehat{y}_i - \widehat{y}_{(i)}}{\sqrt{\widehat{\sigma}^2_{(i)} h_{ii}}}.$$

We consider a point i to be worthy of examination if

$$|DFFITS_i| > 2\sqrt{\frac{p}{n}}$$

Influence Diagnostics (cont.)

- `covratio` change in precision of estimation when ith point is deleted, obtained by considering ratios of determinants; for $i = 1, \ldots, n$,

$$COVRATIO_i = \frac{|\widehat{\sigma}^2_{(i)}(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)})^{-1}|}{|\widehat{\sigma}^2(\mathbf{X}^\top \mathbf{X})^{-1}|} = \left(\frac{\widehat{\sigma}^2_{(i)}}{\widehat{\sigma}^2}\right)^p \frac{1}{1 - h_{ii}}$$

  as

$$\frac{|(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)})^{-1}|}{|(\mathbf{X}^\top \mathbf{X})^{-1}|} = \frac{1}{1 - h_{ii}}$$

- `cooks.distance`: standardized aggregate change in coefficient when ith point is deleted; for $i = 1, \ldots, n$,

$$D_i = \frac{(\widehat{\beta}_{(i)} - \widehat{\beta})^\top (\mathbf{X}^\top \mathbf{X})(\widehat{\beta}_{(i)} - \widehat{\beta})}{p\widehat{\sigma}^2} = \frac{(\widehat{\mathbf{y}}_{(i)} - \widehat{\mathbf{y}})^\top (\widehat{\mathbf{y}}_{(i)} - \widehat{\mathbf{y}})}{p\widehat{\sigma}^2}$$

Adapting the least squares procedure

We may generalize the variance assumption of the linear model to allow for more general variance structures. Suppose that we assume

$$\mathbb{V}\text{ar}[\boldsymbol{\epsilon}|\mathbf{X}] = \sigma^2\mathbf{V}$$

for some positive definite matrix $\mathbf{V}$. This allows the residual errors to have unequal variances and be correlated. Under this assumption, the least squares criterion is amended to

$$\widehat{\beta}_{\mathbf{V}} = \arg\min_{\beta}(\mathbf{y} - \mathbf{X}\beta)^\top\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta)$$

Again we can minimize using calculus to obtain

$$\widehat{\beta}_{\mathbf{V}} = (\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{y}$$

Adapting the least squares procedure (cont.)

Under correct specification of the conditional mean, this estimator is unbiased, with variance

$$\sigma^2(\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{X})^{-1}.$$

We also have

$$\widehat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{y} = \mathbf{H}_\mathbf{V}\mathbf{y}$$

$$\mathrm{SS}_{\mathrm{Res}} = (\mathbf{y} - \widehat{\mathbf{y}})^\top \mathbf{V}^{-1}(\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y}^\top(\mathbf{I}_n - \mathbf{H}_\mathbf{V})\mathbf{V}^{-1}(\mathbf{I}_n - \mathbf{H}_\mathbf{V})\mathbf{y}$$

For convenience, we may write these formulae using

$$\mathbf{W} = \mathbf{V}^{-1}$$

Typically, $\mathbf{V}$ is treated as known, or estimated in a preliminary calculation.

Adapting the least squares procedure (cont.)

We may rewrite this *generalized least squares* formulation as an *ordinary least squares* problem using a decomposition of $\mathbf{V}$ as

$$\mathbf{V} = \mathbf{K}\mathbf{K}^\top = \mathbf{K}^\top\mathbf{K}$$

and premultiplying through the model equation by $\mathbf{K}^{-1}$, that is

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$$

becomes

$$\mathbf{K}^{-1}\mathbf{Y} = \mathbf{K}^{-1}\mathbf{X}\beta + \mathbf{K}^{-1}\boldsymbol{\epsilon}$$

where the variance of $\mathbf{K}^{-1}\boldsymbol{\epsilon}$ is $\sigma^2\mathbf{I}_n$.

Adapting the least squares procedure (cont.)

If **V** is diagonal, with diagonal elements $(v_1, \ldots, v_n)$, then a model can be fitted using weighted least squares via the `lm` function: we let $w = (1/v_1, \ldots, 1/v_n)$, and write

```
lm(y ~ x, weights = w)
```

NB: Using weighted least squares will only affect the **variance** of the estimators.

Regression with transformations

In forming the linear regression model, we may consider transformations of the **predictors** to form part of the conditional expectation model: for example

- polynomial terms: x_{i1}^k, some $k = 1, 2, \ldots$;
- fractional polynomial terms x_{i1}^α, $\alpha \in \mathbb{R}$; eg $\sqrt{x_{i1}}$
- reciprocal terms: $1/x_{i1}$;
- logarithmic terms: $\log x_{i1}$;
- splines, wavelets, orthogonal bases etc.

These can be readily incorporated into the regression model.

Regression with transformations (cont.)

We may also consider transforming the response variable in order to make the standard modelling assumptions

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\beta \qquad \mathbb{V}\text{ar}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}] = \sigma^2\mathbf{I}_n$$

more plausible. For example, we might make the

- $\sqrt{y_i}$;
- $\log y_i$

transformations; typically these transformations are considered *variance stabilizing* transforms.

After the transform, the linear model and constant variance assumptions may appear more plausible.

Regression with transformations (cont.)

For strictly positive responses, the **Box-Cox** family of transformations defines the new response variable as

$$\frac{y_i^\lambda - 1}{\lambda \dot{y}^{\lambda-1}}$$

for some $\lambda \in \mathbb{R}$, where

$$\dot{y} = \exp\left\{\frac{1}{n}\sum_{i=1}^{n}\log y_i\right\}$$

is the geometric mean of $y_1, \ldots, y_n$.

For completeness, we define the new response value when $\lambda = 0$ in this family as

$$\dot{y}\log y_i$$

Example

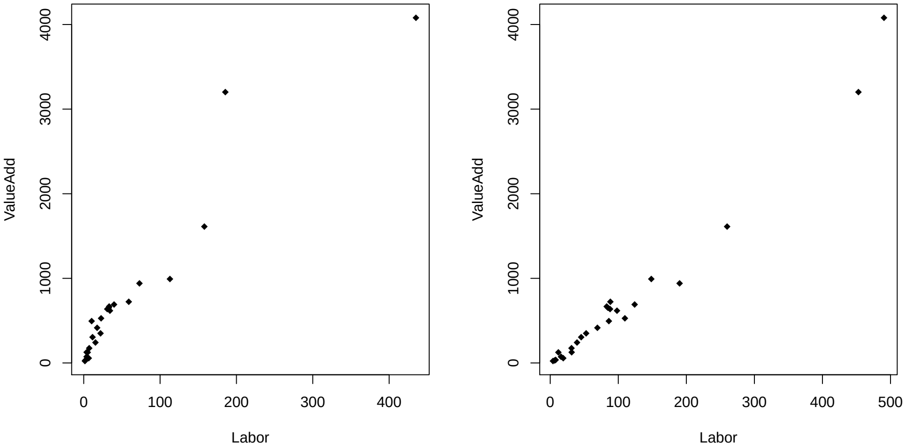
Data on transportation equipment manufacturing: response is value added due to capital investment and other economic indicators. $n = 25$ observations, economic indicators in 1957 dollars.

- `US State` = Observation,
- `ValueAdd` = output (dollars),
- `Capital` = capital input (dollars),
- `Labor` = labour input (dollars),
- `Nfirm` = number of firms.

```
> head(ZD)
        State ValueAdd Capital   Labor NFirm
1     Alabama  126.148   3.804  31.551    68
2  California 3201.486 185.446 452.844  1372
3 Connecticut  690.670  39.712 124.074   154
4     Florida   56.296   6.547  19.181   292
5     Georgia  304.531  11.530  45.534    71
6    Illinois  723.028  58.987  88.391   275
```

Data

Value Added vs Capital and Labour input.

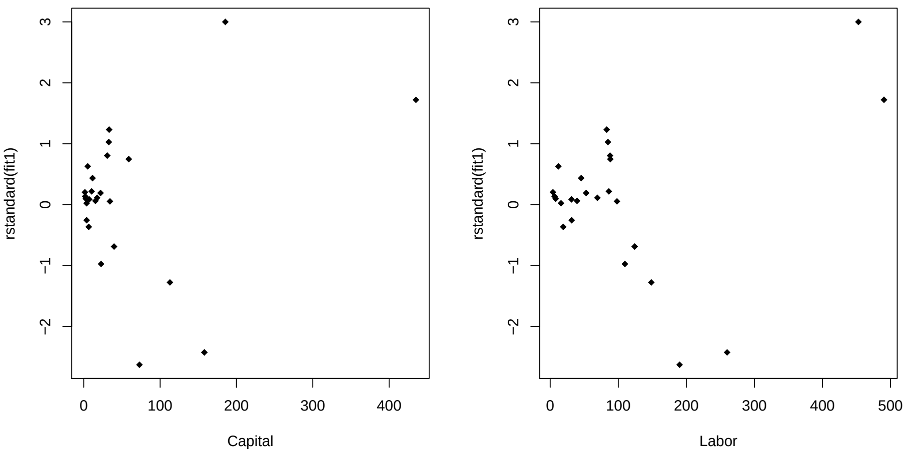


Linear model fit

```
> fit1<-lm(ValueAdd~Capital+Labor,data=ZD)
> summary(fit1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -27.0555    32.9097  -0.822     0.42
Capital       3.2164     0.6461   4.978 5.55e-05 ***
Labor         5.3788     0.4771  11.274 1.31e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 116.4 on 22 degrees of freedom
Multiple R-squared:  0.9867,    Adjusted R-squared:  0.9855
F-statistic:   815 on 2 and 22 DF,  p-value: < 2.2e-16
```

Residuals from linear model fit



Box-Cox transformation

To find the optimal Box-Cox transformation, we need to estimate parameter λ.

This can be done using the `boxcox` function from the MASS library in `R`;

- the model formula is proposed;
- the `boxcox` uses the result of the `lm` to optimize λ.

The `boxcox` function returns a likelihood plot for λ from which the estimate is obtained. This plot is used to construct a confidence interval for λ.

Optimizing λ

For any linear model, the log likelihood for λ is

$$\ell(\lambda; \mathbf{y}, \mathbf{X}) = -\frac{n}{2} \log \mathrm{SS}_{\mathrm{Res}}(\lambda)$$

where $\mathrm{SS}_{\mathrm{Res}}(\lambda)$ is the sum of squared residuals of the given a specified conditional mean model evaluated at a fixed transformation value λ.

By standard maximum likelihood theory, we have that an approximate 95 % interval for λ is the collection of values of t such that

$$2n(\ell(\widehat{\lambda}; \mathbf{Y}, \mathbf{X}) - \ell(t; \mathbf{Y}, \mathbf{X})) \leq 3.841$$

where 3.841 is the 0.95 quantile of the χ^2_1 distribution.

Box-Cox Analysis

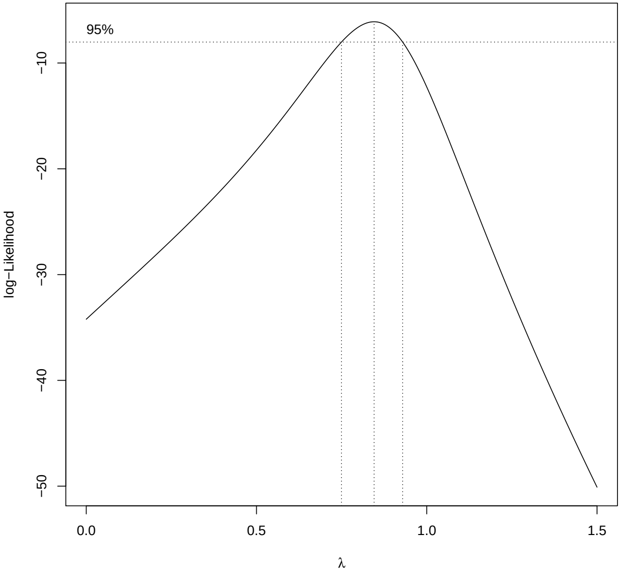
```
> lam.fit<-boxcox(lm(ValueAdd~Capital+Labor,data=ZD),
+                 plotit=FALSE,lambda=seq(0,1.5,by=0.0001))
>
> (lambda.hat<-lam.fit$x[which.max(lam.fit$y)])
[1] 0.8453
```

Hence $\widehat{\lambda} = 0.8453$. The call

```
lam.fit<-boxcox(lm(ValueAdd~Capital+Labor,data=ZD),
+                lambda=seq(0,1.5,by=0.0001))
```

automatically produces a plot.

Box-Cox log-likelihood plot

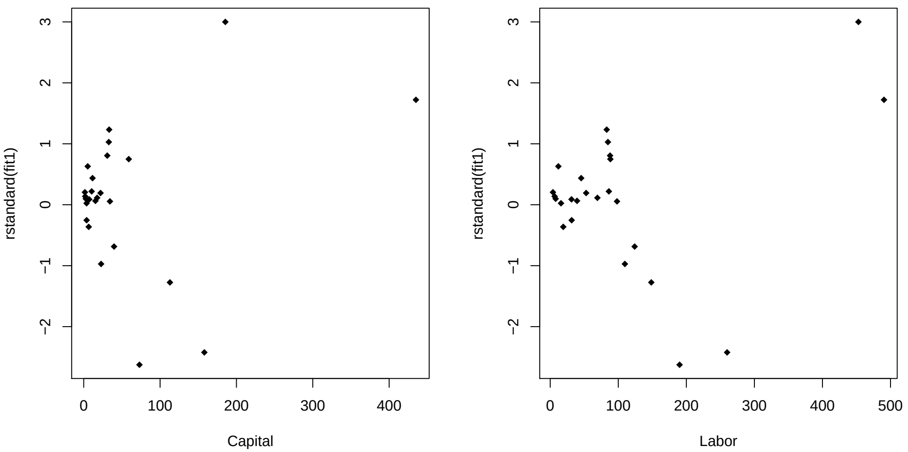


Transforming the response

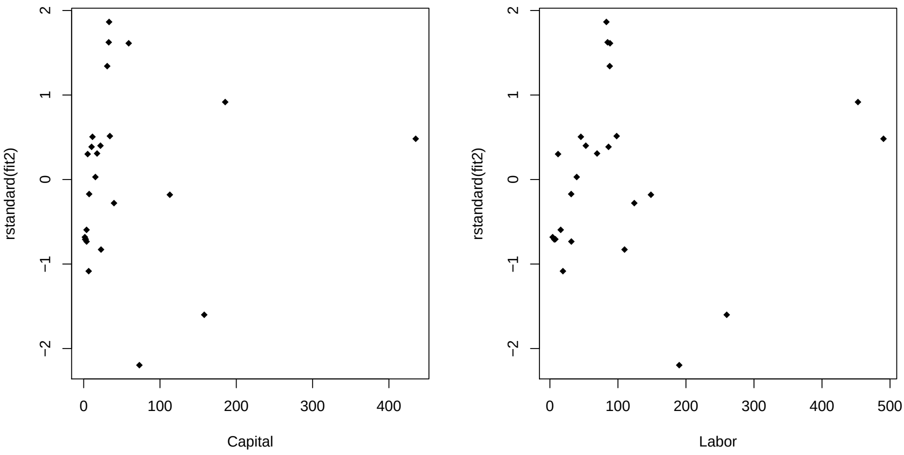
```
> ytilde<-exp(mean(log(ZD$ValueAdd)))
> Y<-ZD$ValueAdd
> ZD$ynew<-(Y^lambda.hat-1)/(lambda.hat*ytilde^(lambda.hat-1))
> fit2<-lm(ynew~Capital+Labor,data=ZD)
> summary(fit2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  74.7825    25.7096   2.909  0.00814 **
Capital       1.8260     0.5047   3.618  0.00152 **
Labor         4.8738     0.3727  13.076 7.51e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 90.97 on 22 degrees of freedom
Multiple R-squared:  0.9875,    Adjusted R-squared:  0.9863
F-statistic: 866.1 on 2 and 22 DF,  p-value: < 2.2e-16
```

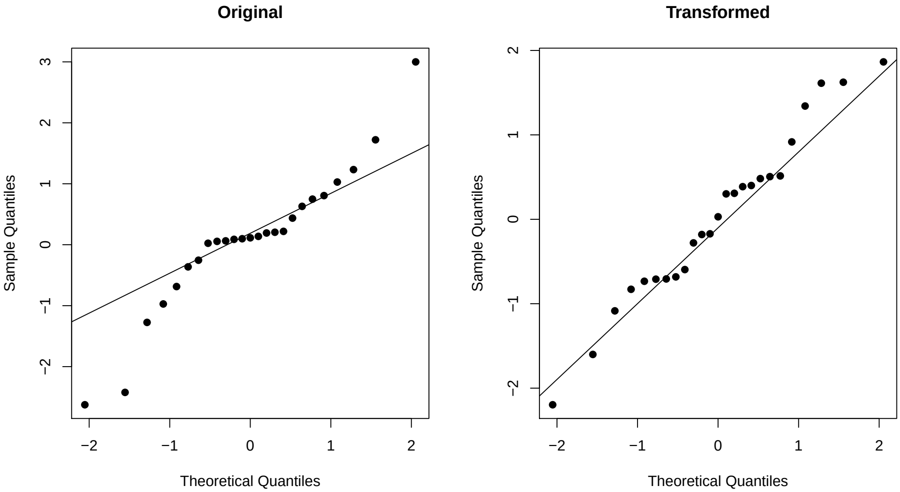
Residuals from linear model fit



Residuals from transformed model fit



Normal Q-Q plots



Summary

- Transformed data are more suitable for linear regression modelling with the fit computed by least squares;
- Optimal λ here is `0.8453`;
- R^2 values are similarly high in the two cases, but the results are not directly comparable due the data transformation.

Cobb-Douglas Production Function

The Cobb-Douglas production function for observed economic data $i = 1, \ldots, n$ may be expressed as

$$O_i = e^{\beta_0} l_i^{\beta_1} c_i^{\beta_2} u_i$$

where

- O_i is output
- l_i is labour input
- c_i is capital input
- u_i is a random error term

Cobb-Douglas Production Function (cont.)

Taking natural logs, we have that

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

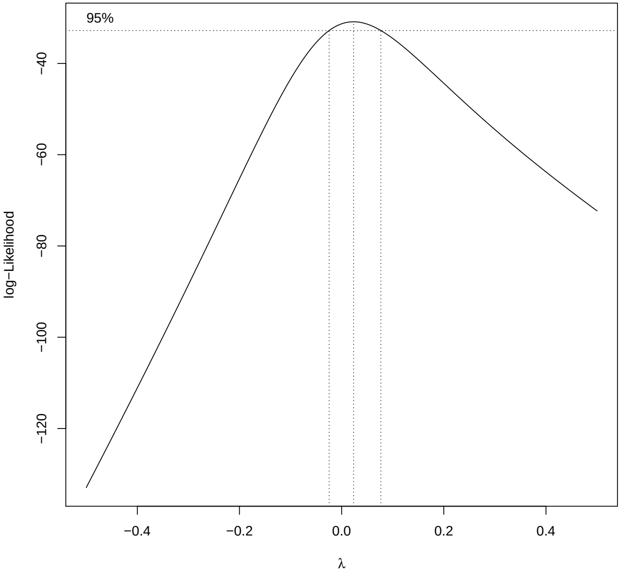
where

- $Y_i = \ln(O_i)$ is log output
- $x_{i1} = \ln(l_i)$ is log labour input
- $x_{i2} = \ln(c_i)$ is log capital input
- $\epsilon_i = \ln(u_i)$ is a random error term

Cobb-Douglas Production Function

```
>CD<-read.csv(’CobbDouglas.csv’)
>y<-(as.numeric(gsub(",","",as.character(CD$Y))))
>x1<-log(as.numeric(gsub(",","",as.character(CD$X2))))
>x2<-log(as.numeric(gsub(",","",as.character(CD$X3))))
>boxcox(lm(y ~ x1+x2),lambda=seq(-0.5,0.5,by=0.0001))
```

Box-Cox log-likelihood plot



Box-Cox log-likelihood plot

The conclusion from the log-likelihood plot is that λ is not significantly different from zero.

Therefore the log transform (which corresponds to $\lambda = 0$) is appropriate.

Appendix

Computing the PRESS residuals efficiently

We have

$$e_{(i)} = y_i - \widehat{y}_{(i)} = y_i - \mathbf{x}_i \widehat{\beta}_{(i)} = y_i - \mathbf{x}_i (\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)})^{-1} \mathbf{X}_{(i)}^\top \mathbf{y}_{(i)}.$$

Now

$$\mathbf{X}_{(i)}^\top \mathbf{y}_{(i)} = \mathbf{X}^\top \mathbf{y} - \mathbf{x}_i^\top y_i$$

and

$$\mathbf{X}^\top \mathbf{X} = \sum_{j=1}^{n} \mathbf{x}_j^\top \mathbf{x}_j = \mathbf{x}_i^\top \mathbf{x}_i + \sum_{j \neq i} \mathbf{x}_j^\top \mathbf{x}_j = \mathbf{x}_i^\top \mathbf{x}_i + \mathbf{X}_{(i)}^\top \mathbf{X}_{(i)}.$$

so therefore

$$(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)})^{-1} = \left(\mathbf{X}^\top \mathbf{X} - \mathbf{x}_i^\top \mathbf{x}_i \right)^{-1}$$

Woodbury's Matrix Formula

We have that

$$(\mathbf{A}+\mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1}+\mathbf{DC}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}.$$

where

- $\mathbf{A}$ is $p \times p$;
- $\mathbf{B}$ is $p \times q$;
- $\mathbf{C}$ is $q \times q$;
- $\mathbf{D}$ is $q \times p$;

Woodbury's Matrix Formula (cont.)

We set

- $\mathbf{A} = \mathbf{X}^\top\mathbf{X}$ is $p \times p$;
- $\mathbf{B} = -\mathbf{x}_i^\top$ is $p \times 1$;
- $\mathbf{C} = 1$ is 1×1;
- $\mathbf{D} = \mathbf{x}_i$ is $1 \times p$;

so that

$$\left(\mathbf{X}^\top\mathbf{X} - \mathbf{x}_i^\top\mathbf{x}_i\right)^{-1} = (\mathbf{X}^\top\mathbf{X})^{-1} + \frac{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i^\top\mathbf{x}_i(\mathbf{X}^\top\mathbf{X})^{-1}}{1 - \mathbf{x}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i^\top}$$

where

$$1 - \mathbf{x}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i^\top = 1 - h_{ii}$$

Woodbury's Matrix Formula (cont.)

Therefore

$$y_i - \mathbf{x}_i(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}^\top \mathbf{y}_{(i)}$$

simplifies to

$$y_i - \mathbf{x}_i\left((\mathbf{X}^\top\mathbf{X})^{-1} + \frac{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i^\top \mathbf{x}_i(\mathbf{X}^\top\mathbf{X})^{-1}}{1-h_{ii}}\right)\left(\mathbf{X}^\top\mathbf{y} - \mathbf{x}_i^\top y_i\right)$$

where

$$\mathbf{x}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{x}_i\widehat{\beta} = \widehat{y}_i$$

This becomes

$$y_i - \mathbf{x}_i \left(\widehat{\beta} + \frac{1}{1 - h_{ii}} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i^\top \mathbf{x}_i \widehat{\beta} \right)$$

$$+ \mathbf{x}_i \left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i^\top y_i + \frac{1}{1 - h_{ii}} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i^\top \mathbf{x}_i (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i^\top y_i \right)$$

or equivalently

$$y_i - \widehat{y}_i - \frac{h_{ii}}{1 - h_{ii}} \widehat{y}_i + h_{ii} y_i + \frac{h_{ii}^2}{1 - h_{ii}} y_i$$

which simplifies to

$$\frac{y_i - \widehat{y}_i}{1 - h_{ii}}.$$

Assessing Normality via Probability Plots

For a collection of residuals $e_i, i = 1, \ldots, n$, we may check whether the Normality assumption is violated using probability plotting.

- **P-P plot**: plot
  - x-axis: the values

$$\frac{i - 1/2}{n} \qquad i = 1, \ldots, n$$

  - y-axis: if the residuals are sorted into ascending order.

$$e_1 < e_2 < \cdots < e_n.$$

    we plot

$$\Phi\left(\frac{e_i - \bar{e}}{s_e}\right).$$

    where $\Phi(.)$ is the standard normal cdf, $\bar{e}$ is the sample mean of the residuals, and s_e is the sample standard deviation of the residuals.

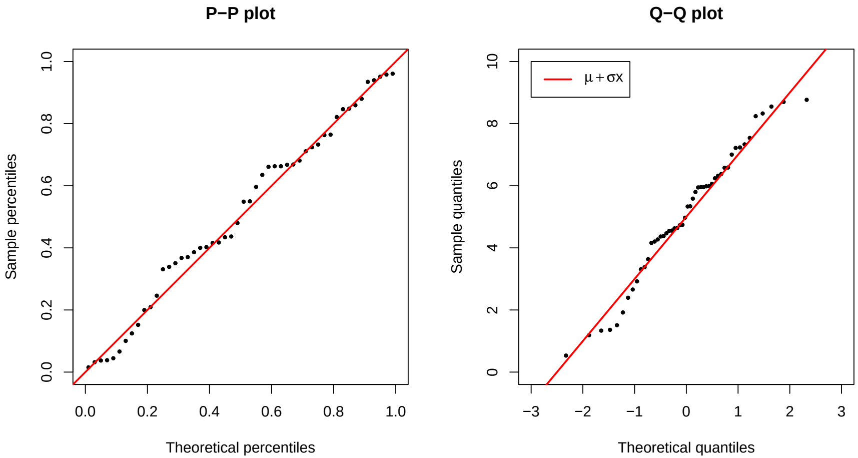
Assessing Normality via Probability Plots (cont.)

- **Q-Q plot**
  - x-axis: the values

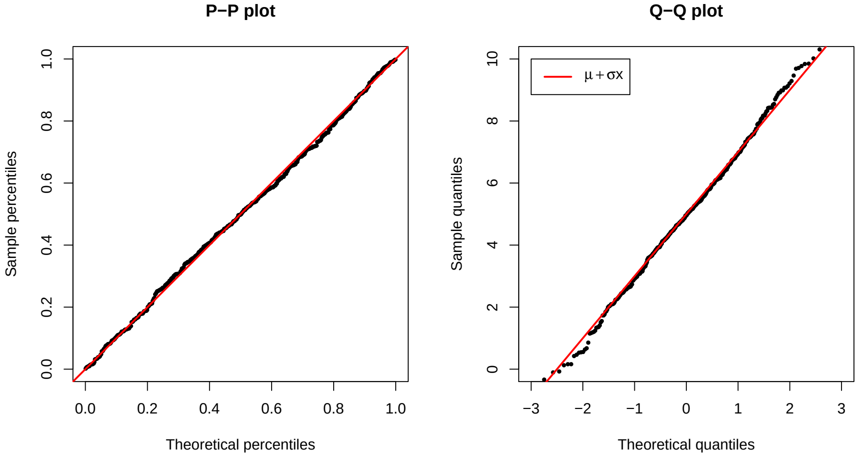
    $$\Phi^{-1}\left(\frac{i-1/2}{n}\right) \qquad i = 1, \ldots, n$$

    where $\Phi^{-1}(.)$ is the standard normal inverse cdf;
  - y-axis: the residuals sorted into ascending order.

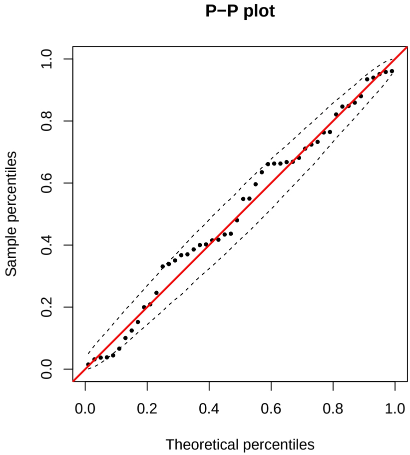
P-P/Q-Q plots: $n = 50$

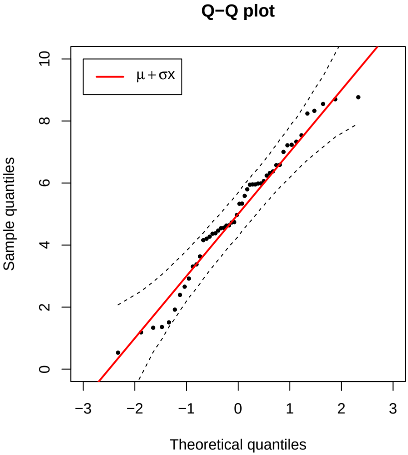


P-P/Q-Q plots: $n = 500$



P-P/Q-Q plots with 95 % CI: $n = 50$





P-P/Q-Q plots with 95 % CI: $n = 500$