

Diagnostics in Multiple Linear Regression

University of California, San Diego

Instructor: Armin Schwartzman

<http://math.ucsd.edu/~eariasca/teaching.html>

1 / 25

Checking assumptions

- All the p-values for the tests we performed and the confidence levels for all the confidence intervals we computed were derived under the standard assumptions.

- **Standard assumptions.** The measurement errors are **i.i.d. normal with mean zero**:

$$\varepsilon_1, \dots, \varepsilon_n \sim^{iid} \mathcal{N}(0, \sigma^2)$$

- Are those assumptions correct?

1. Mean zero, i.e. model accuracy \rightarrow unbiased
2. Equal variances, i.e. homoscedasticity \rightarrow believe estimate of covariance
3. Independence
4. Normality \rightarrow normality of estimate

2 / 25

Residuals are not independent \rightarrow core to each other

- We (obviously) do not have access to the errors. Instead, we look at the residuals as proxies.

- Remember that $\mathbf{e} = (e_1, \dots, e_n)$, with

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

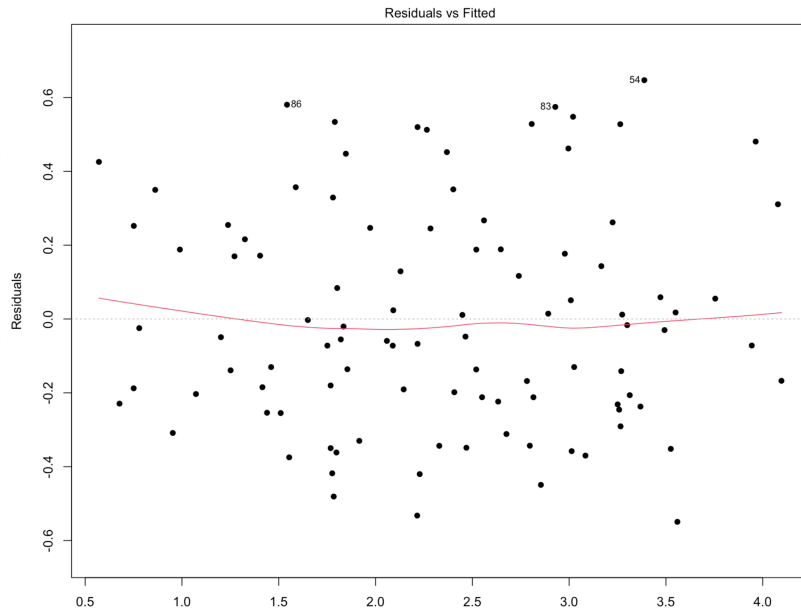
- Under the standard assumptions, \mathbf{e} is **multivariate normal** with mean zero and covariance $\sigma^2(\mathbf{I} - \mathbf{H})$.

- In particular,

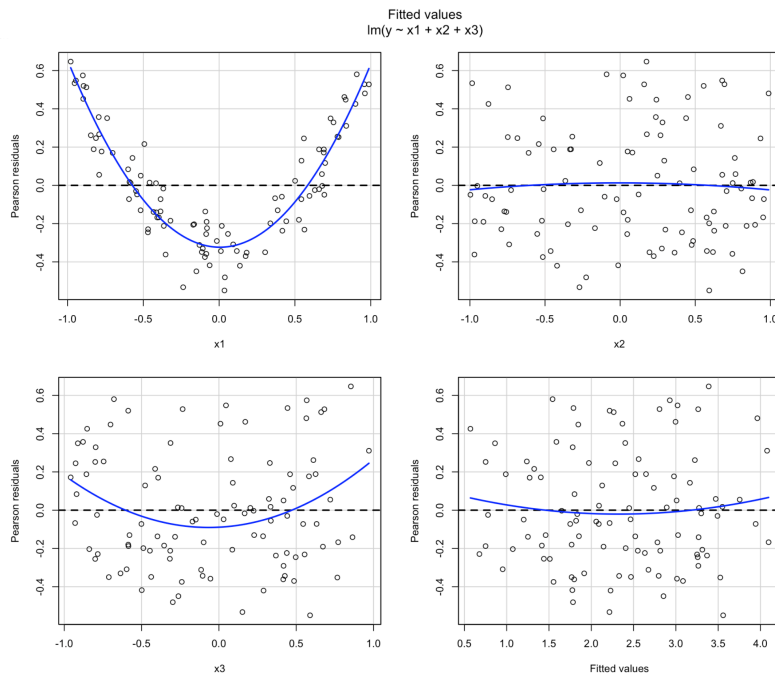
$$\text{Cov}(e_i, e_j) = -\sigma^2 h_{ij} \quad \forall i \neq j, \quad \text{Var}(e_i) = \sigma^2(1 - h_{ii}) \quad \forall i$$

3 / 25

$$\begin{aligned} \text{cov}(\mathbf{e} | \mathbf{X}) &= \text{cov}[(\mathbf{I} - \mathbf{H})\mathbf{y} | \mathbf{X}] \\ &= (\mathbf{I} - \mathbf{H}) \text{cov}(\mathbf{y} | \mathbf{X}) (\mathbf{I} - \mathbf{H})^T \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= (\mathbf{I} - \mathbf{H}) \text{cov}(\boldsymbol{\varepsilon}) (\mathbf{I} - \mathbf{H})^T \\ &= (\mathbf{I} - \mathbf{H}) \sigma^2 \mathbf{I} \cdot (\mathbf{I} - \mathbf{H})^T \\ &= \sigma^2 (\mathbf{I} - \mathbf{H}) (\mathbf{I} - \mathbf{H})^T \\ &= \sigma^2 (\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) (\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) = \sigma^2 (\mathbf{I} - \mathbf{H}) \end{aligned}$$



non parametric
mean estimator



→ linear x_2 x_3

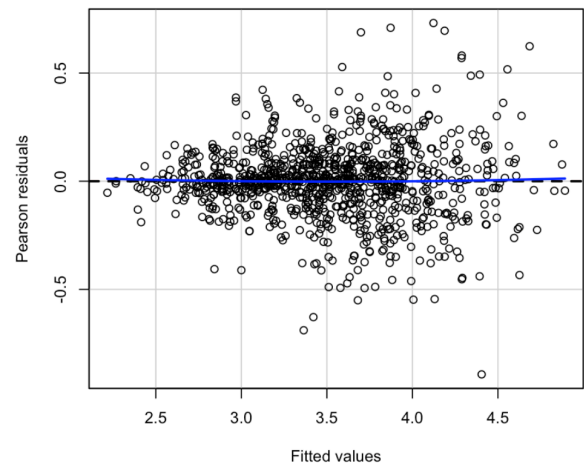
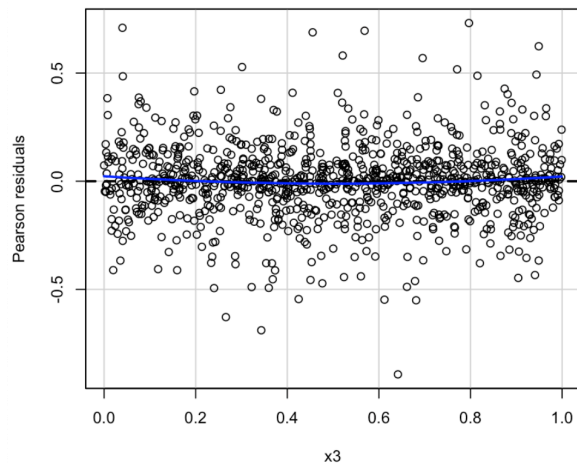
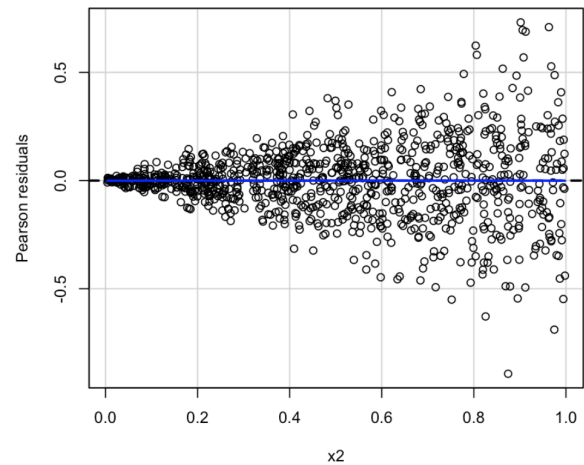
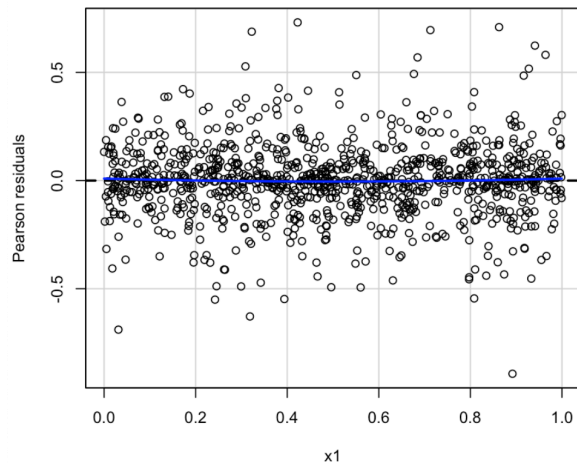
quadratic x_1

```
3 # an example of inaccurate model
4 x1 = runif(100,-1,1)
5 x2 = runif(100,-1,1)
6 x3 = runif(100,-1,1)
7 y = 2 + x1^2 + x2 + x3 + rnorm(100,sd=.1)
8 ex.lm = lm(y ~ x1 + x2 + x3) # we are not fitting the correct model
9 summary(ex.lm)
10
11 pairs(y~ x1 + x2 + x3, pch=16, cex=1.5) # only a hint of curvature on the scatterplots
12 plot(ex.lm, which=1, pch=16) # the residuals vs fitted values plot does not show much, so we look at partial residual plots
13 residualPlots(ex.lm) # Residual plots
14 # all the partial residual plots show a clear curvature in x1
```

```

16 # an example of heteroscedastic model
17 x1 = runif(1000)
18 x2 = runif(1000)
19 x3 = runif(1000)
20 y = 2 + x1 + x2 + x3 + 3*x2*rnorm(1000, sd=.1)
21 ex.lm = lm(y ~ x1 + x2 + x3)
22 pairs(y ~ x1 + x2 + x3) # not much is visible
23 residualPlots(ex.lm) # clear fan shape on the 2nd plot

```



```

Residuals:
    Min       1Q   Median       3Q      Max
-0.89298 -0.07754  0.00159  0.08127  0.73119

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.01525    0.01760   114.51  <2e-16 ***
x1           0.98444    0.01913    51.47  <2e-16 ***
x2           1.00581    0.01961    51.28  <2e-16 ***
x3           0.98508    0.01908    51.62  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1732 on 996 degrees of freedom
Multiple R-squared:  0.8892,    Adjusted R-squared:  0.8888
F-statistic: 2664 on 3 and 996 DF, p-value: < 2.2e-16

```

Standardized residuals

- In practice, the residuals are often corrected for unequal variance.
- **Internally studentized residuals** (what R uses):

$$r_i = \frac{e_i}{\hat{\sigma} \sqrt{1 - h_{ii}}}$$

- Note that: $r_i \overset{\text{approximation}}{\sim} T_{n-p-1}$

- **Externally studentized residuals:**

$$t_i = \frac{e_i}{\hat{\sigma}_{(i)} \sqrt{1 - h_{ii}}}$$

where $\hat{\sigma}_{(i)}$ comes from the fit with the i th observation removed.

- Note that: $t_i \sim T_{n-p-2}$

4 / 25

Checking model accuracy

- With several predictors, the residuals-vs-fitted-values plot (provided by R) may be misleading. Instead, we look at **partial residual plots**, which focus on one variable at a time.
- Say we want to look at the influence of predictor $\mathbf{X}_j = (x_{1,j}, \dots, x_{n,j})$ (the j th column of \mathbf{X}) on the response \mathbf{y} .
- **Partial residual plots:** \mathbf{e} versus \mathbf{X}_j .
- **Component plus residual plots:** $\hat{\beta}_j \mathbf{X}_j + \mathbf{e}$ versus \mathbf{X}_j .

They allow to better appreciate the variation of the residuals compared to the variation in each component.

- **Added variable plots** take into account the influence of the other predictors:

1. Regress \mathbf{y} on all the predictors excluding \mathbf{X}_j , yielding residuals $\mathbf{y}_{(j)}$.
2. Regress \mathbf{X}_j on all the predictors excluding \mathbf{X}_j , yielding residuals $\mathbf{X}_{(j)}$.
3. Plot $\mathbf{y}_{(j)}$ versus $\mathbf{X}_{(j)}$.

(For each method, the procedure is repeated for all $j = 1, \dots, p$.)

5 / 25

Checking homoscedasticity

- ☐ The residuals vs fitted values plot is what R provides by default and can be helpful in situations where σ varies with $\mathbb{E}(y|\mathbf{x})$.
- ☐ Partial residual plots may be helpful when σ depends on one variable only.
- ☐ A fan-shape in the plot is an indication that the errors do not have equal variances.
- ☐ Testing for equality of variances is **not recommended**. There are methods (e.g., Levene test) for that but they are somewhat sensitive to non-normality. Moreover, the tests and confidence intervals we computed are somewhat robust to mild heteroscedasticity.

6 / 25

Checking normality

- ☐ R provides a **q-q plot** of the standardized residuals.
- ☐ Testing for normality is **not recommended**. There are methods (e.g., Lilliefors test) for that but the tests and confidence intervals we computed are somewhat robust to mild departures from normality.

7 / 25

Checking for independence

- ☐ Checking for independence often requires defining a dependency structure explicitly and testing against that. (For otherwise we would need a number of replicates of the data, i.e. multiple samples of same size.)
- ☐ If the order of the observations is not arbitrary, for example the observations are collected over time, then this might introduce some dependency. Such **serial dependency** may be tested for (e.g., the **Durbin-Watson** test).

If serial correlation is expected, then it is preferable to model the errors using some model, e.g., an **autoregressive model** of some fixed order $q \geq 1$. In that case, a likelihood approach is still viable, but the computations are more complicated (and do not result in closed-form expressions).

8 / 25

Outliers

- ☐ **Outliers** are points that are unusual compared to the bulk of the observations.
- ☐ An **outlier in predictor** (aka **high-leverage point**) is a data point (\mathbf{x}_i, y_i) such that \mathbf{x}_i is away from the bulk of the sample predictor vectors.
- ☐ An **outlier in response** is a data point (\mathbf{x}_i, y_i) such that y_i is away from the trend implied by the other observations.
- ☐ A point that is both an outlier in predictor and response is said to be **influential**.

9 / 25

Detecting outliers

- To detect **outliers in predictor**, plotting the **hat values** h_{ii} may be useful. The hat values are the diagonal entries of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$.
- Rule of thumb: $h_{ii} > 2(p+1)/n$ is considered problematic.
- Suppose (without loss of generality) that the variables have been normalized so that $\bar{\mathbf{x}} = 0$ and there is no intercept in the fit. Then

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

It measures the distance of \mathbf{x}_i from $\bar{\mathbf{x}} = 0$ in the (Mahalanobis) metric given by $\mathbf{X}^\top \mathbf{X}$.

- To detect **outliers in response**, plotting the **externally studentized residuals** may be useful. We know they are normalized to have the same T_{n-p-2} distribution under 'normal' circumstances.

10 / 25

Leave-one-out diagnostics

- If $\hat{\boldsymbol{\beta}}_{(i)}$ is the least squares coefficient vector computed with the i th case deleted, then

$$\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - \frac{e_i}{1 - h_{ii}} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i$$

- Define

$$\hat{\mathbf{y}}_{(i)} = \mathbf{X} \hat{\boldsymbol{\beta}}_{(i)}$$

In particular, $\hat{y}_{(i)i}$ is the value at \mathbf{x}_i predicted by the model fitted without the observation (\mathbf{x}_i, y_i) . Note that

$$\hat{y}_{(i)i} = \hat{y}_i - \frac{e_i}{1 - h_{ii}} \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i$$

11 / 25

Cook's distances

- **Cook's distances**

$$\begin{aligned} D_i &= \frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)}\|^2}{(p+1)\hat{\sigma}^2} \quad (\text{change in fitted values}) \\ &= \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^\top \mathbf{X}^\top \mathbf{X} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{(p+1)\hat{\sigma}^2} \quad (\text{change in coefficients}) \\ &= \frac{r_i^2}{p+1} \frac{h_{ii}}{1 - h_{ii}} \quad (\text{combination of residuals and hat values}) \end{aligned}$$

- Rule of thumb: $D_i > 1$ is considered suspect.

12 / 25

DFBETAS and DFFITS

□ DFBETAS

$$\text{DFBETAS}_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{\hat{\sigma}_{(i)}^2 (\mathbf{X}^\top \mathbf{X})_{j,j}^{-1}}}$$

- Rule of thumb: $\text{DFBETAS}_{j(i)} > 2/\sqrt{n}$ is considered suspect.

□ DFFITS

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{\hat{\sigma}_{(i)}^2 h_{ii}}} = t_i \sqrt{\frac{h_i}{1 - h_i}}$$

- Rule of thumb: $\text{DFFITS}_i > 2\sqrt{(p+1)/n}$ is considered suspect.

13 / 25

Multicollinearity

- When the predictors are nearly linearly dependent, several issues arise:

1. Interpretation is difficult.
2. The estimates have **large variances**:

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{(1 - Q_j^2) \text{SS}_{X_j}}$$

where Q_j^2 is the R^2 when regressing \mathbf{X}_j on $\mathbf{X}_k, k \neq j$.

3. The fit is **numerically unstable** since $\mathbf{X}^\top \mathbf{X}$ is almost singular.
- Assume we work with standardized variables, so that there is no intercept and the predictor vectors \mathbf{X}_j have zero mean and unit variance (use the function scale). (The columns can be standardized to unit norm, which corresponds to a variance equal to $1/(n-1)$.)

14 / 25

Correlation between predictors

- A large correlation between \mathbf{X}_j and \mathbf{X}_k is indicative that these variables are nearly linearly dependent (in this case, proportional).
- We simply inspect the correlation between the variables. Note that with standardized variables, the correlation between \mathbf{X}_j and \mathbf{X}_k is simply their inner product $\mathbf{X}_j^\top \mathbf{X}_k$. In particular, $\mathbf{X}^\top \mathbf{X}$ is the correlation matrix in this case.

15 / 25

$$\begin{aligned}
 y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon && x_1, x_2 \text{ are dependent} \\
 &= \beta_0 + \beta_1 x_1 + \beta_2 (\gamma_0 + \gamma_1 x_1 + \delta) + \varepsilon \\
 &= (\beta_0 + \beta_2 \gamma_0) + (\beta_1 + \beta_2 \gamma_1) x_1 + (\beta_2 \delta + \varepsilon)
 \end{aligned}$$

Suppose $\delta = 0$:

$$y = (\beta_0 + \beta_2 \gamma_0) + (\beta_1 + \beta_2 \gamma_1) x_1 + \varepsilon$$

Variance Inflation Factors

- Alternatively, one can look at the **variance inflation factors**:

$$\text{VIF}_j = \frac{1}{1 - Q_j^2}$$

- If the predictors are standardized to unit variance, $\mathbf{X}^\top \mathbf{X}$ is a correlation matrix and we have:

$$\text{VIF} = \text{diag}((\mathbf{X}^\top \mathbf{X})^{-1})$$

$\text{VIF}_j > 10$ (same as $Q_j^2 > 90\%$) is considered suspect

16 / 25

Condition Indices

- Another option is to examine the **condition indices** of $\mathbf{X}^\top \mathbf{X}$:

$$\kappa_j = \frac{\lambda_{\max}}{\lambda_j}$$

where the λ_j 's are the eigenvalues of $\mathbf{X}^\top \mathbf{X}$.

$\kappa_j > 1000$ is considered suspect

(It is important to standardize the variables for this to make sense.)

- $\lambda_{\max}/\lambda_{\min}$ is an important quantity in numerical linear algebra. It is called the **condition number** of matrix $\mathbf{X}^\top \mathbf{X}$ and quantifies how stable it is to invert a linear system of the form

$$(\mathbf{X}^\top \mathbf{X})\mathbf{b} = \mathbf{c}$$

(Alternatively, one can use the singular values of \mathbf{X} , which are the square root of the eigenvalues of $\mathbf{X}^\top \mathbf{X}$.)

17 / 25

Dealing with curvature

- When the model accuracy is questionable, we can:

1. **Transform** the variables and/or the response, which often involves guessing. This is particularly useful to **spread** the variables — which helps limit the number of high-leverage observations.
2. **Augment** the model by adding other terms, perhaps with interactions. This is often used in conjunction with **model selection** so as to avoid overfitting.

18 / 25

Dealing with heteroscedasticity

□ When homoscedasticity is questionable, we can:

1. Apply a **variance stabilizing transformation** to the response.
2. Fit the model by **weighted least squares**.

Both involve guessing at the dependency of the variance as a function of the variables.

19 / 25

Variance stabilizing transformations

□ Here are a few common transformations on the response to stabilize the variance:

| | |
|--|---------------------------------|
| $\sigma^2 \propto \mathbb{E}(y_i)$ (Poisson) | change to \sqrt{y} |
| $\sigma^2 \propto \mathbb{E}(y_i)(1 - \mathbb{E}(y_i))$ (binomial) | change to $\sin^{-1}(\sqrt{y})$ |
| $\sigma^2 \propto \mathbb{E}(y_i)^2$ | change to $\log(y)$ |
| $\sigma^2 \propto \mathbb{E}(y_i)^3$ | change to $1/\sqrt{y}$ |
| $\sigma^2 \propto \mathbb{E}(y_i)^4$ | change to $1/y$ |

□ In general, we want a transformation ψ such that

$$\psi'(\mathbb{E}(y_i)) \text{Var}(y_i) \propto 1$$

(This is based on the **delta method**, which involves asymptotic calculations.)

20 / 25

Weighted least squares

□ Suppose

$$y_i = \beta^\top \mathbf{x}_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

□ Maximum likelihood estimation of β corresponds to the weighted least squares solution that minimizes

$$\text{SSE}(\beta) = \sum_{i=1}^n w_i (y_i - \beta^\top \mathbf{x}_i)^2, \quad w_i = \frac{1}{\sigma_i^2}$$

□ To determine appropriate weights, different approaches are used:

- ▷ Guess how σ_i^2 varies with x_i , i.e. the shape of $\text{Var}(y_i|x_i)$.
- ▷ Assume a parametric model for the weights and use maximum likelihood estimation – solved by **iteratively reweighted least squares**.

□ Using weights is helpful in situations where some observations are more reliable than others, mostly because of variability.

21 / 25

Generalized least squares

- ☐ Weighted least squares assumes that the errors are uncorrelated.
- ☐ **Generalized least squares** assumes a more general form for the covariance of the errors, namely $\sigma^2\mathbf{V}$, where \mathbf{V} is usually known.
- ☐ Assuming both the design matrix \mathbf{X} and the covariance matrix \mathbf{V} are full rank, the maximum likelihood estimates are

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

22 / 25

Dealing with outliers

- ☐ **Spreading** the variables often equalizes the leverage of the data points. It often improves the fit, both numerically (e.g. R^2) and visually.
- ☐ A typical situation is an accumulation of data points near 0, in which case applying a logarithm or a square root helps spread the data more evenly.
- ☐ Example: `mammals` dataset in the `MASS` package.

23 / 25

Dealing with outliers

- ☐ We have identified an outlier. Was it recorded properly?
 - ▷ No: correct it or remove it from the fit.
 - ▷ Yes: decide whether to include it in the fit.
- ☐ Genuine outliers carry information, so simply discarding them amounts to losing that information. However, strong outliers can compromise the quality of the overall fit.
- ☐ Possible options:
 1. If there are comparatively few of them, remove the outliers (particularly the influential points).
 2. If there are many outliers, model outliers and non-outliers separately. There might be a lurking variable that we may need to include in the model.
 3. Use a robust method for fitting the model.

24 / 25

Dealing with Multicollinearity

- ☐ If **interpretation** is the main goal, drop a few redundant predictors.
- ☐ If **prediction** is the main goal, use a **model selection** procedure.

25 / 25