Regression Assumptions and Diagnostics

Max Turgeon

Department of Statistics, University of Manitoba

Learning Outcomes

- Understand how regression relates to statistical inference.
- Recognize the relative importance of regression assumptions.
- Be able to assess evidence that an assumption is likely not met, and how to refine a regression model accordingly.

Statistical Inference

Gelman, Hill and Vehtari (2020) describe the three main challenges of statistical inference:

1. Generalizing from *sample* to *population*.
2. Generalizing from *treatment* to *control* group.
3. Generalizing from observed *measurements* to the underlying *constructs of interest*.

Regression and Inference

- **Regression** allows us to study how *average* values of an *outcome* variable vary across individuals. Each individual is defined by a set of *covariates*.
- Applications:
    - Prediction
    - Exploring associations
    - Adjusting for confounders
    - Causal inference

Lifecycle of a regression model

1. Model building
2. Model fitting
3. **Understanding the fit**
4. Criticism

Linear Regression

Recall: Linear model

- Y is an outcome variable, $X_1, \ldots, X_p$ are covariates.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \mathrm{error}.$$

- Here, error is a random variable with mean 0 and variance σ^2, so we can also write

$$E(Y \mid X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

- The coefficients β_i represent comparisons of **means** for different values of the covariates (i.e. for different individuals).

Assumptions

Gelman, Hill and Vehtari (2020) list the assumptions of linear regression **in decreasing order of importance**:

1. Validity (with respect to the research question).
2. Representativeness (of the data with respect to the population).
3. Additivity and linearity.
4. Independence of errors.
5. Equal variance of errors.
6. Normality of errors.

Validity and Representativeness i

- The most important assumptions of linear regression are **non-mathematical**.
    - They are entirely based on domain knowledge
- **Validity**
    - Outcome measure should reflect question of interest
    - Relevant predictors/risk factors should be included
    - Model should generalize to patients to which results will be applied
- **Representativeness**
    - Data can be used to make inference about a larger population.
    - Including more covariates into model can help bridge the “representativeness” gap between data and population.

Validity and Representativeness ii

- How to fix this? The solution is often to change the model.
    - *Validity*: Measurement error models.
    - *Representativeness*: IPT weights, selection models.
    - When all else fails, you may have to narrow the scope of your research question (e.g. more descriptive than causal).

Additivity and linearity

- Main mathematical assumption:

$$E(Y \mid X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

- Or in English:
    - Changes in the conditional mean of Y should be additive and linear.
- **Note**: Conditional mean = on average
    - Life is probably nonlinear and non-additive…
    - But it can still be a good approximation of the average

Diagnostic plots

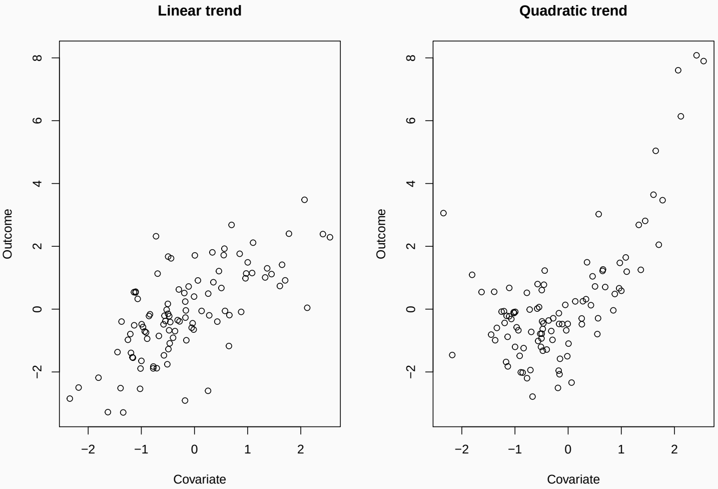
1. For **simple** linear regression (i.e. only one covariate), plot outcome against covariate.
2. Plot outcome against fitted values.
3. Plot residuals against fitted values and/or covariates.

Note: It is not recommended to plot outcome against residuals.

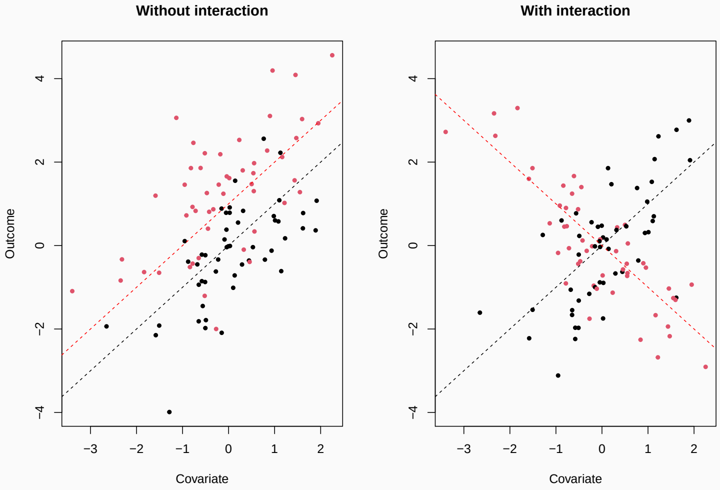
Outcome vs Covariate i

- This is the simplest case, because we can actually *visualize the relationship*.
- But it *only* works with a single covariate.
    - Or two, if one is categorical
- We are looking for evidence that we could fit a line through point cloud.
    - Or perhaps we need to fit a quadratic term, etc.

Outcome vs Covariate ii



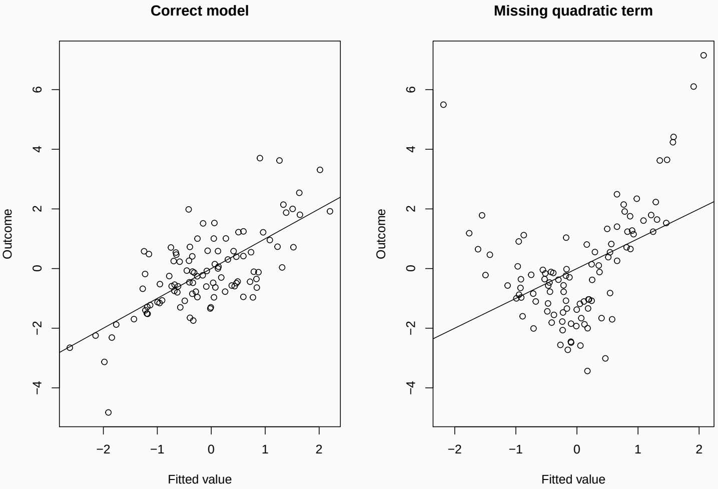
Outcome vs Covariate iii



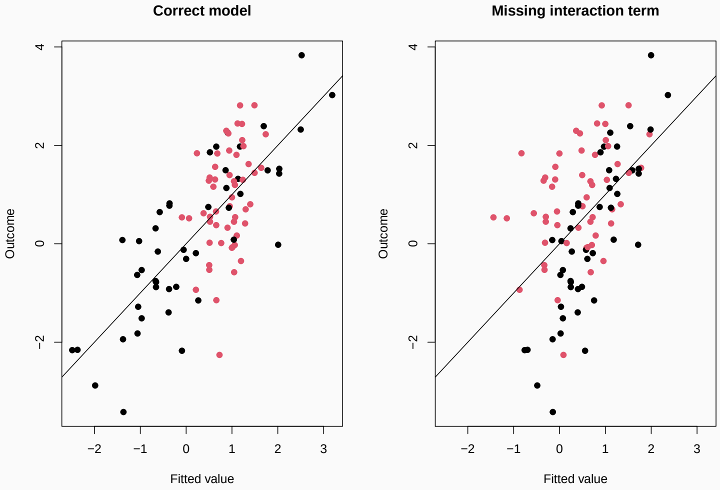
Outcome vs Fitted values i

- *Recall*: the fitted values are *estimates* of the conditional mean of the outcome
    - Outcome variable should be randomly distributed around its conditional mean.
- Therefore, we expect outcome vs. fitted should follow diagonal.
    - Otherwise, part of the variation is not explained (e.g. because of missing covariate).

Outcome vs Fitted values ii



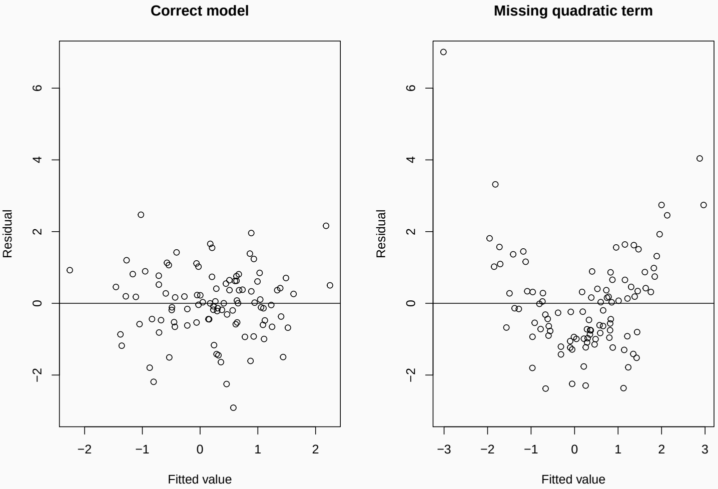
Outcome vs Fitted values iii



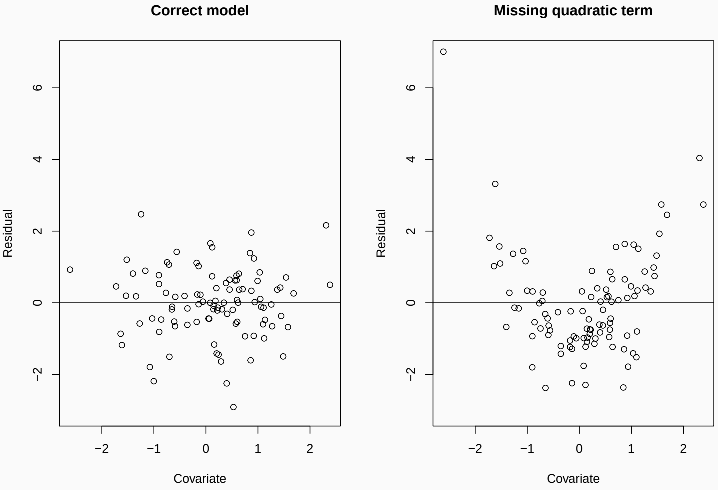
Residuals vs Fitted values/Covariate i

- *Recall*: the residuals are the difference between the outcome and the fitted values.
  - They should be independent of covariate and fitted values
- Therefore, we expect residuals vs fitted values/covariate to follow a horizontal line.
  - Otherwise, part of the variation is not explained (e.g. because of missing covariate).

Residuals vs Fitted values/Covariate ii



Residuals vs Fitted values/Covariate iii



How to fix this?

- Transform outcome variable.
  - Eg. Using logarithms, multiplicative effects become additive.
  - **Note**: It changes interpretation of regression coefficients.
- Transform covariates.
  - **Note**: It changes interpretation of regression coefficients.
- Add quadratic term or splines to model nonlinear trends.
  - **Note**: If the same variable appears in multiple terms (e.g. linear and quadratic term), you can no longer vary one while keeping the other fixed.
- Add interaction term.
  - **Note**: It changes interpretation of regression coefficients.

Example i

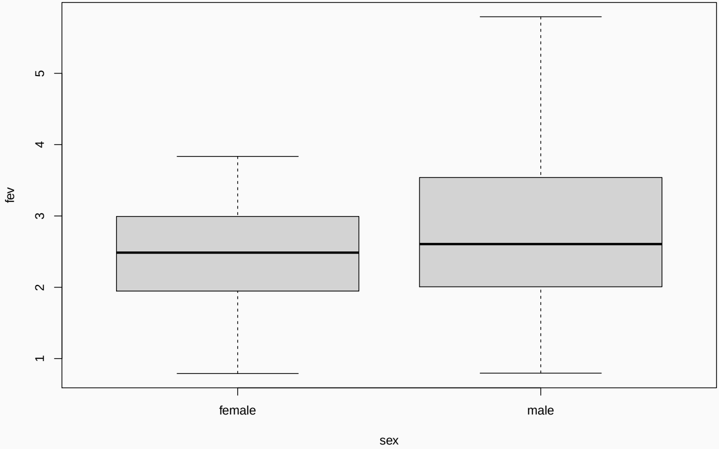
- We will use data on Forced Expiratory Volume (FEV) in children age 3 to 19 from East Boston recorded during the 1970s.
    - Can be downloaded from `http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets`
- The dataset contains information on age, height, sex, and smoking status.
- **Outcome**: FEV

Example ii

```
# Import dataset into R
data_fev <- read.csv(”FEV.csv”)

# Explore data
boxplot(fev ~ sex, data = data_fev)
```

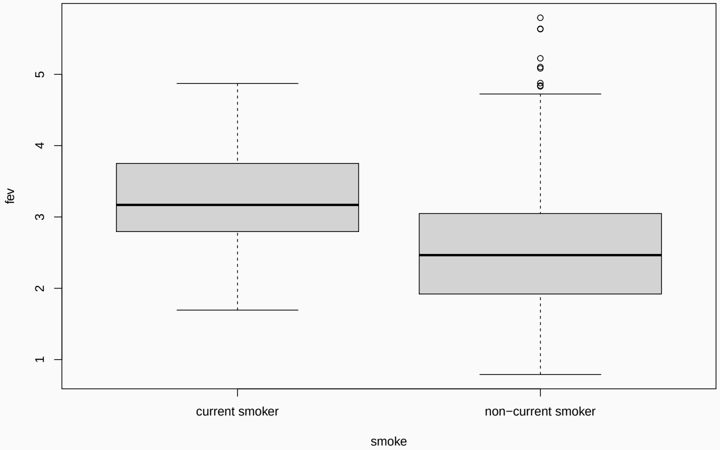
Example iii



Example iv

```r
boxplot(fev ~ smoke, data = data_fev)
```

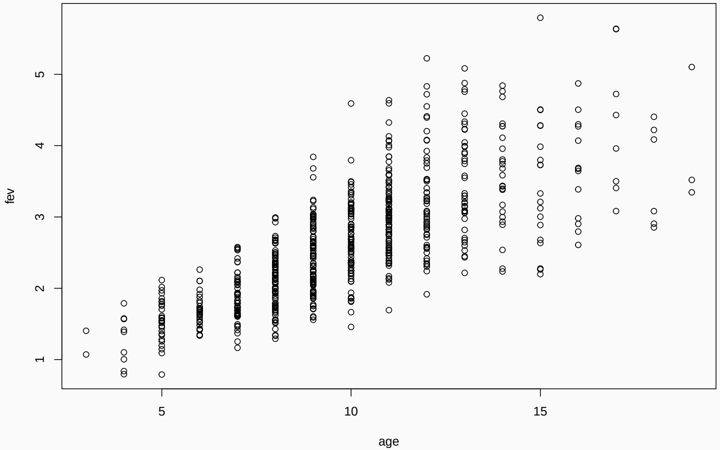
Example v



Example vi

```r
# Note: Use 'with' instead of 'attach'
with(data_fev, plot(age, fev))
```

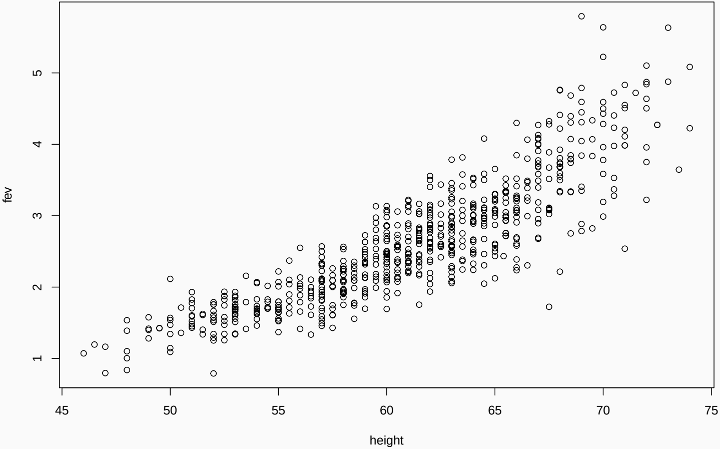
Example vii



Example viii

```r
with(data_fev, plot(height, fev))
```

Example ix



Example x

```r
# Fit linear model
model <- lm(fev ~ smoke + sex + age + height,
            data = data_fev)
```

Example xi

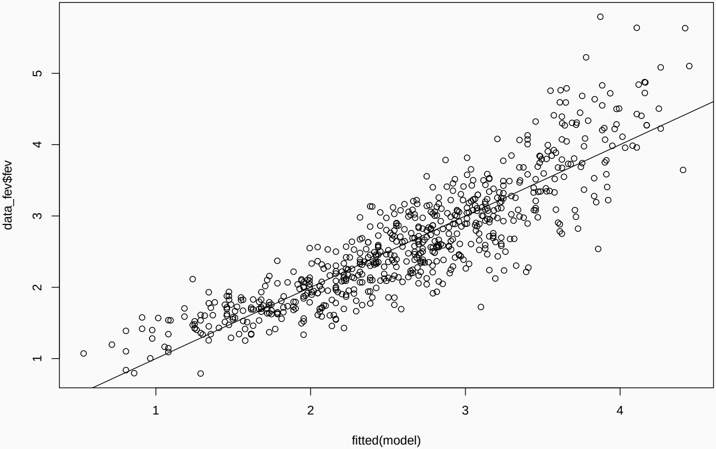
```r
# Output nice table
knitr::kable(broom::tidy(model),
             digits = 2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-4.54	0.23	-19.58	0.00
smokenon-current smoker	0.09	0.06	1.47	0.14
sexmale	0.16	0.03	4.73	0.00
age	0.07	0.01	6.90	0.00
height	0.10	0.00	21.90	0.00

Example xii

```r
# Plot outcome vs fitted values
plot(fitted(model), data_fev$fev)
# Add diagonal line
abline(a = 0, b = 1)
```

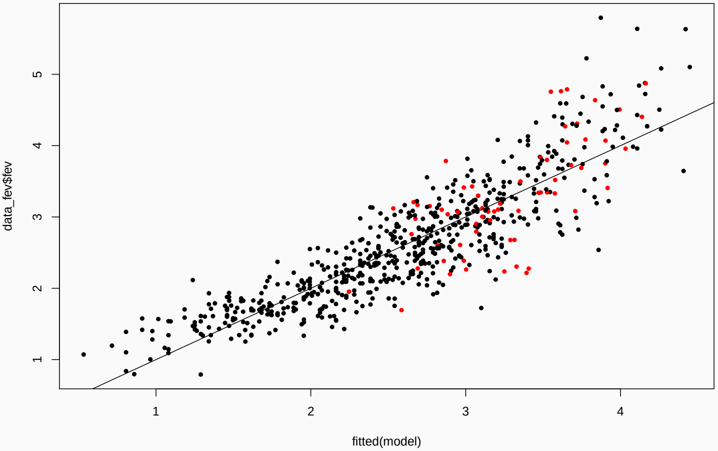
Example xiii



Example xiv

```
# Can also colour points according to smoking status
colour <- ifelse(data_fev$smoke == ”current smoker”,
                 ”red”, ”black”)
plot(fitted(model), data_fev$fev,
     col = colour, pch = 20)
# Add diagonal line
abline(a = 0, b = 1)
```

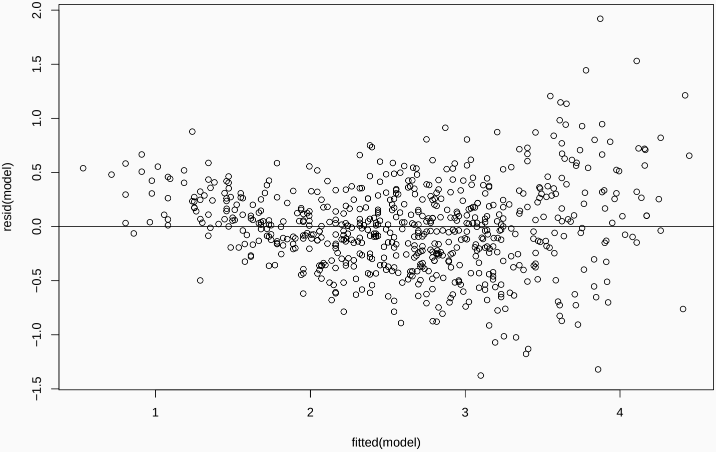
Example xv



Example xvi

```r
# Plot residuals vs fitted values
plot(fitted(model), resid(model))
# Add horizontal line
abline(h = 0)
```

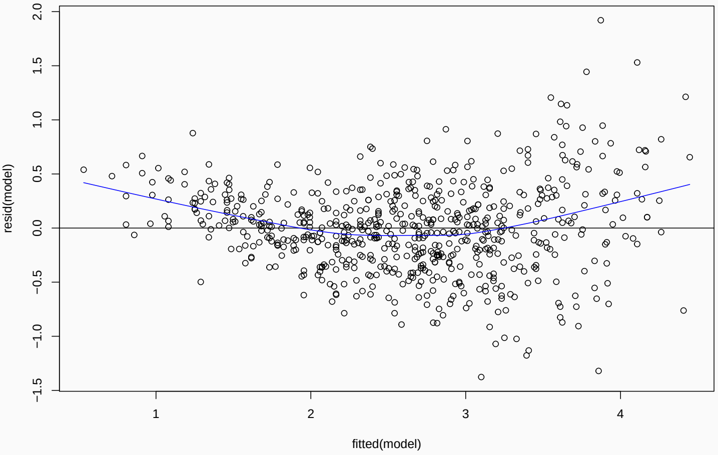
Example xvii



Example xviii

```
# Can also add a nonlinear smoother
plot(fitted(model), resid(model))
abline(h = 0)
# We will use LOWESS
lines(lowess(fitted(model), resid(model)),
      col = ”blue”)
```

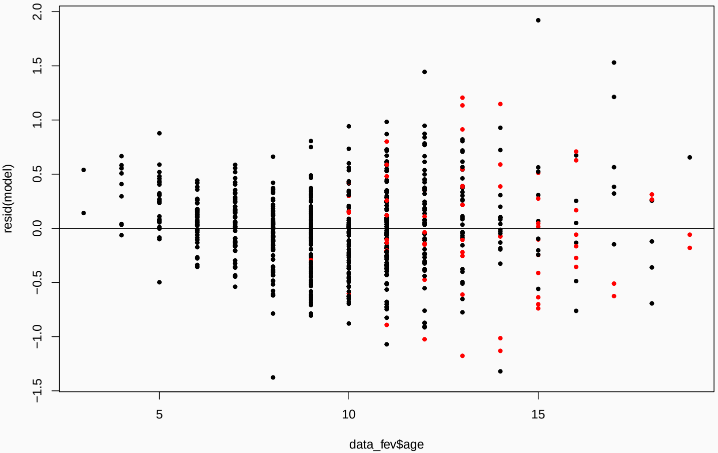
Example xix



Example xx

```r
# Plot residuals vs age
plot(data_fev$age, resid(model),
     col = colour, pch = 20)
abline(h = 0)
```

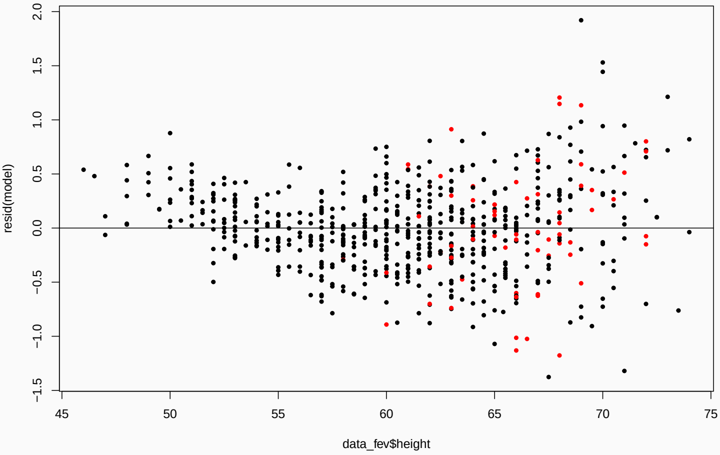
Example xxi



Example xxii

```r
# Plot residuals vs height
plot(data_fev$height, resid(model),
     col = colour, pch = 20)
abline(h = 0)
```

Example xxiii



Interpretation of coefficients i

- Transformations change the interpretation of the coefficients.
- Let's say we have a linear regression model where the outcome is earnings (in 1000$) and the covariates are height (in inches) and sex:

$$\text{Earnings} \sim -26 + 0.6\text{Height} + 10.6\text{Male}.$$

- *Interpretation*: On average, a person one inch taller than another person of the same sex earns 600$ more.
- Alternative model: the outcome is earnings (in 1000$) and the covariates are **logarithm** of height and sex:

$$\text{Earnings} \sim -162 + 42.7\text{logHeight} + 10.7\text{Male}.$$

Interpretation of coefficients ii

- *Interpretation*: We need to do a bit of algebra.

$$E(\text{Earnings} \mid \text{Height} = x) = -162 + 42.7 \log(x)$$
$$E(\text{Earnings} \mid \text{Height} = x + 1) = -162 + 42.7 \log(x + 1)$$

- Therefore, the difference in average earnings when a person is one inch taller is given by

$$\begin{aligned} &(-162 + 42.7 \log(x + 1)) - (-162 + 42.7 \log(x)) \\ &\quad = 42.7 \log \left(\frac{x + 1}{x} \right). \end{aligned}$$

- In particular, it depends on x, the baseline height!

Interpretation of coefficients iii

- This is why we instead use *multiplicative* changes when using covariates on the logarithmic scale.
- Let's compare two people, with one 10% taller than the other one:

$$
\begin{aligned}
E(\text{Earnings} \mid \text{Height} = x) &= -162 + 42.7 \log(x) \\
E(\text{Earnings} \mid \text{Height} = 1.1x) &= -162 + 42.7 \log(1.1x)
\end{aligned}
$$

- Therefore, the difference in average earnings when a person is 10% taller is about 4070$:

$$
\begin{aligned}
&(-162 + 42.7 \log(1.1x)) - (-162 + 42.7 \log(x)) \\
&\quad = 42.7 \log\left(\frac{1.1x}{x}\right) = 42.7 \log\left(1.1\right) \approx 4070\text{\$}.
\end{aligned}
$$

Interpretation of coefficients iv

- Alternative model: the outcome is earnings (on the **log scale**) and the covariates are height and sex:

$$\log - \text{Earnings} \sim 8.0 + 0.02\text{Height} + 0.4\text{Male}.$$

- *Interpretation*: On average, a person one inch taller than another person of the same sex earns 0.02 log-dollars more.
- We would like to interpret this on the original scale, but the logarithm of the average is **not** equal to the average of the values on the logarithmic scale.
- **However**, if log-earnings are approximately symmetric, we know that mean = median.

Interpretation of coefficients v

  - And the median is preserved under the logarithm!
  - *Remember*: Difference on the log scale is a ratio on the original scale.
- *Interpretation 2*: Since $\exp(0.02) = 1.02$, the median income of a person one inch taller than another person of the same sex is 2% higher.

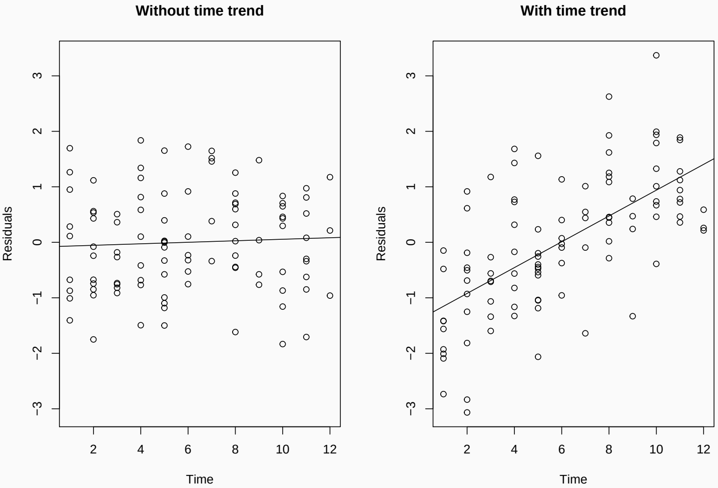
Independence of errors

- Independence of the errors is important when performing hypothesis testing and calculating confidence intervals.
  - With dependent data, tests are too optimistic and CIs are too narrow.
- On the other hand, the effect on the coefficient estimates should be minimal.
- **When is it not met?** The main source of dependent data is *clustered* or grouped (e.g. patients in a hospital, weather sensors in a province).
- **How to fix this?** Use mixed models or generalized estimating equations. Or add clustering variable into the model.

Diagnostic i

- Dependence between the errors is usually driven by time dependence (i.e. order in which the observations were taken), spatial dependence, or clustering.
- Diagnostic: Plot residuals against any of these variables. Departure from a horizontal trend is evidence of correlation.
    - *Tip*: Use boxplots when clustering variable is discrete.

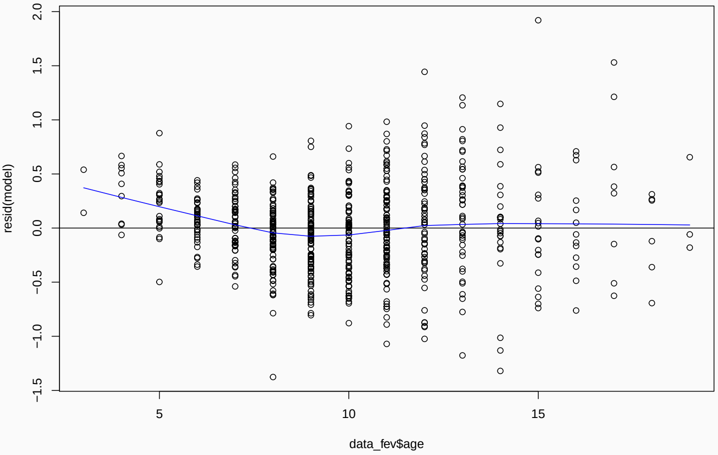
Diagnostic ii



Example i

```r
# Residuals vs age
plot(data_fev$age, resid(model))
abline(h = 0)
lines(lowess(data_fev$age, resid(model)),
      col = ”blue”)
```

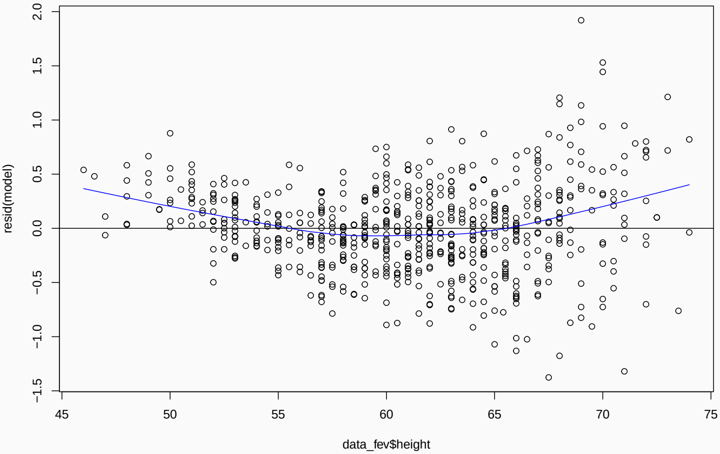
Example ii



Example iii

```r
# Residuals vs height
plot(data_fev$height, resid(model))
abline(h = 0)
lines(lowess(data_fev$height, resid(model)),
      col = ”blue”)
```

Example iv


resid(model)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
45
50
55
60
65
70
75
data_fev$height

Equal variance of errors i

- Equal variance (aka homoscedasticity) is actually a fairly unimportant assumption.
    - If the goal of the model is prediction, accounting for unequal variance will improve accuracy.
- Unequal variance (aka heteroscedasticity) does not affect the frequentist properties of the inference.
    - Hypothesis tests are valid, and so are the confidence intervals.
- However, accounting for unequal variance can lead to more efficient inference (i.e. lower variance, narrower CIs).

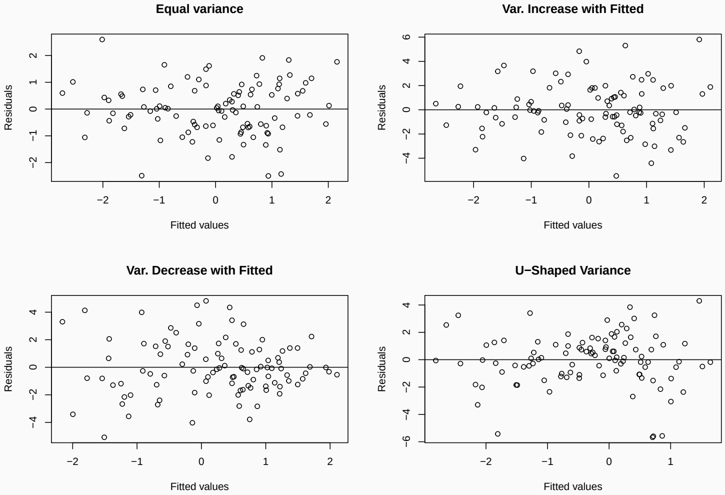
Equal variance of errors ii

- **When is it not met?** Unequal variance could simply be a feature of the data, and it is common to have the variance depend on covariates (e.g. higher income patients have more variability in their diet).
- **How to fix this?** Weighted linear regression or Eicker–Huber–White standard errors.

Diagnostic i

- One way to see evidence of unequal variance is to plot the *residuals* against the *fitted values*.
    - Equal variance means residuals should randomly fall within a band around the horizontal line $y = 0$.
- If there is evidence of heteroscedasticity, you can try to find the source by plotting *residuals* against individual *covariates*.

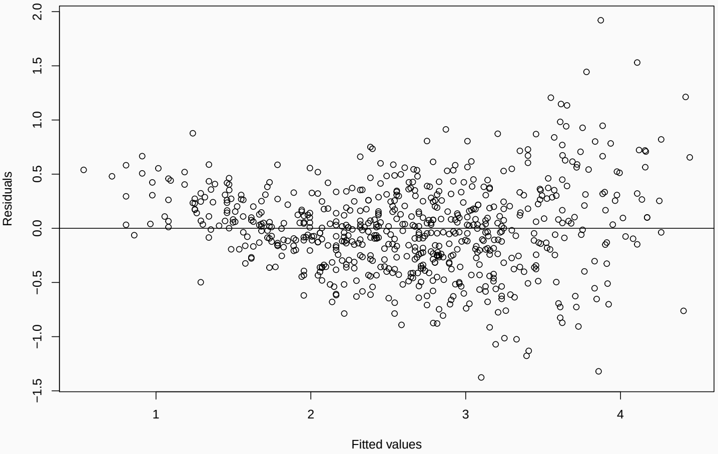
Diagnostic ii



Example i

```r
# Recall: FEV data
model <- lm(fev ~ smoke + sex + age + height,
            data = data_fev)
# Plot residuals vs fitted
plot(fitted(model), resid(model),
     xlab = ”Fitted values”, ylab = ”Residuals”)
abline(h = 0)
```

Example ii



Example iii

- There is evidence of heteroscedasticity.
- Let’s look at the Eicker–Huber–White standard errors.
    - **Note**: For Stata, use the option `robust` of the `regress` procedure.

```
# Default standard errors
knitr::kable(subset(broom::tidy(model),
                    select = c(”term”, ”estimate”,
                               ”std.error”)),
             digits = 3)
```

Example iv

term	estimate	std.error
(Intercept)	-4.544	0.232
smokenon-current smoker	0.087	0.059
sexmale	0.157	0.033
age	0.066	0.009
height	0.104	0.005

Example v

```r
# EHW standard errors
vcov <- sandwich::vcovHC(model)
knitr::kable(cbind(coef(model),
                   sqrt(diag(vcov))),
             digits = 3)
```

(Intercept)	-4.544	0.258
smokenon-current smoker	0.087	0.078
sexmale	0.157	0.032
age	0.066	0.010
height	0.104	0.005

Example vi

	Estim.	Std. CI	Robust CI
(Intercept)	-4.544	(-5, -4.089)	(-5.05, -4.039)
smokenon-current smoker	0.087	(-0.029, 0.204)	(-0.065, 0.24)
sexmale	0.157	(0.092, 0.222)	(0.094, 0.22)
age	0.066	(0.047, 0.084)	(0.045, 0.086)
height	0.104	(0.095, 0.114)	(0.094, 0.114)

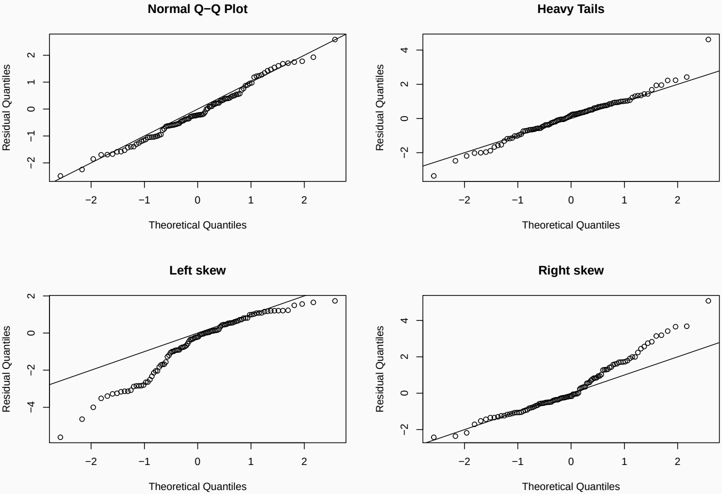
Normality of errors

- Normality of the errors is the least important assumption.
    - Frankly, its purpose is to make the math easier.
- Non-normality is only important for prediction.
    - It does not affect inference.
- **When is it not met?** Pretty much all the time!
- **How to fix this?** Use prediction intervals based on more appropriate distribution (e.g. t distribution).

Diagnostic i

- One way to diagnose non-normality is to look at **QQ-plots**.
    - We know the mean of the residuals is zero, and we can estimate its variance $\widehat{\sigma^2}$.
    - We can compare the quantiles of the residuals with those of a normal distribution $N(0, \widehat{\sigma^2})$.
- It is **not** recommended to test the hypothesis of normality.

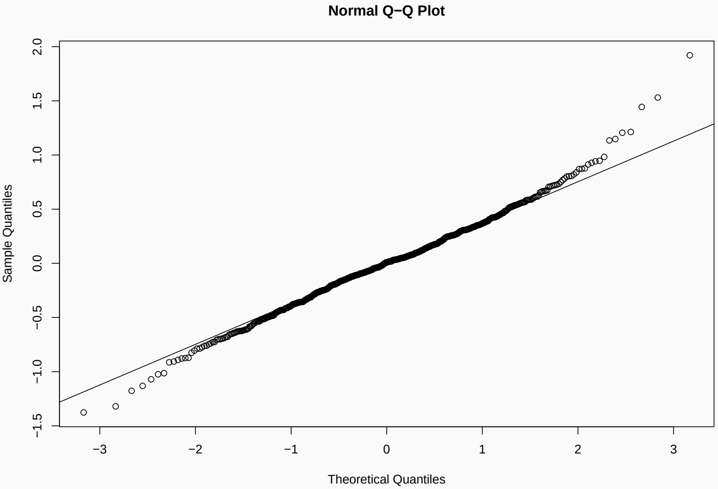
Diagnostic ii



Example i

```r
# Evidence of heavier tails
qqnorm(resid(model))
qqline(resid(model))
```

Example ii



Normal Q−Q Plot
Sample Quantiles
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−3
−2
−1
0
1
2
3
Theoretical Quantiles

Other considerations i

- R^2 measures *how much variation in the outcome variable is explained by the model.*
  - What constitutes a good R^2 is highly dependent on the problem.
  - Not a good metric for assessing model fit.
  - **Never use for model selection** (it is inherently biased towards complex models)
- High correlation between covariates can lead to large standard errors and wide confidence intervals.
  - This is known as *multicollinearity*.
  - It is measured using *kappa* (aka *condition number*) or *variance inflation factor*.

Other considerations ii

    - It can be fixed by removing/combining/transforming some covariates.

- There is plethora of *influence measures* (e.g. Leverage values, Cook's distance).
    - These measures can be helpful to uncover outliers.
    - Understanding why an observation is an outlier can be helpful in refining your model (especially if "being an outlier" is correlated with other variables).
    - However, none of these measures are fail-proof; they are helpful diagnostics.

Example i

```r
# R2 values
summary(model)$r.squared
```

```
## [1] 0.7753614
```

```r
summary(model)$adj.r.squared
```

```
## [1] 0.7739769
```

Example ii

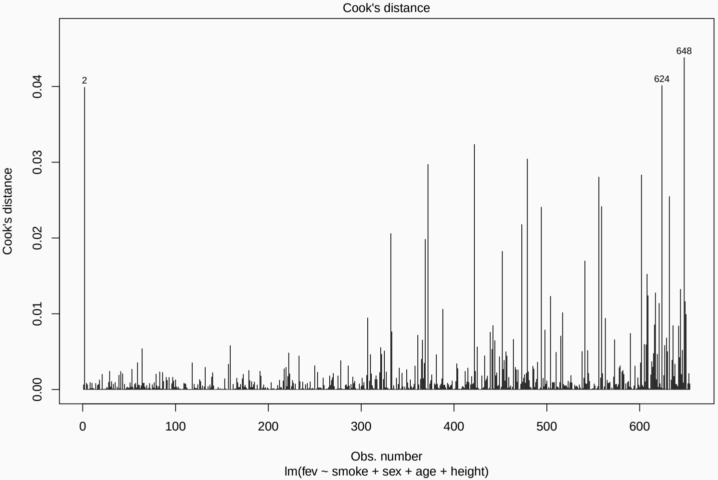
```r
# Evaluate Multicollinearity----
# Variance Inflation Factors
car::vif(model)
```

```
##    smoke      sex      age   height
## 1.209564 1.060228 3.019010 2.829728
```

Example iii

```r
# Cook's D plot----
# identify D values > 4/(n-p-1)
n <- nrow(data_fev)
p <- length(coef(model)) - 1
cutoff <- 4/(n - p - 1)
plot(model, which = 4,
     cook.levels = cutoff)
```

Example iv


Cook's distance
Cook's distance
0.00
0.01
0.02
0.03
0.04
2
624
648
0
100
200
300
400
500
600
Obs. number
lm(fev ~ smoke + sex + age + height)

Example v

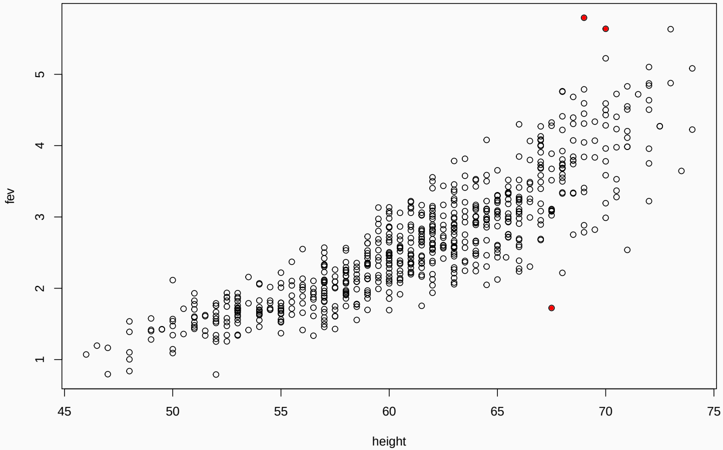
```r
# Look at raw data
data_fev[c(2, 624, 648),]
```

```
##         id age   fev height    sex              smoke
## 2      451   8 1.724   67.5 female non-current smoker
## 624 25941  15 5.793   69.0   male non-current smoker
## 648 71141  17 5.638   70.0   male non-current smoker
```

Example vi

```r
with(data_fev, plot(height, fev))
# Colour outliers in red
with(data_fev[c(2, 624, 648),],
     points(height, fev, col = ”red”, pch = 20))
```

Example vii



Discussion and Summary i

- We found evidence that additivity/linearity is not met.
  - Residual vs fitted plot, but also residual vs height.
  - Given our data visualizations, it is likely that relationship between FEV and height is nonlinear.
  - We could address this using a logarithmic transformation or splines.
- We found evidence of heteroscedasticity.
  - Residual vs fitted values; higher variance with larger fitted values.
  - We computed robust standard errors but saw no major change in our inference.

Discussion and Summary ii

- We found evidence of a few outliers.
    - But after closer look at the raw data, they do not seem like implausible values.
- Model checking is an *iterative process*.
- It is also more an art than a science.
    - In particular, it is easier to find evidence *against* than evidence *for*.
- Diagnostic plots are preferable to hypothesis tests.

Logistic Regression

Recall: Logistic regression

- Y is a binary outcome variable (i.e. $Y = 0$ or $Y = 1$).

$$\operatorname{logit}\left(E(Y \mid X_1, \ldots, X_p)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

- Recall: $\operatorname{logit}(t) = \log(t/(1-t))$.
- The coefficients β_i represent comparisons of log odds for different values of the covariates (i.e. for different individuals).

Assumptions

Logistic regression has less assumptions than linear regression.

1. Validity (with respect to the research question).
2. Representativeness (of the data with respect to the population).
3. Additivity and linearity.
4. (Conditional) Independence of the outcomes.

Note: There is only one possible distribution for binary outcomes, i.e. Bernoulli. As a consequence, we **always** have heteroscedasticity.

Diagnostic plots i

- Diagnostic plots are trickier with logistic regression because the data is *discrete*.
    - And therefore the *residuals* are also discrete.
- One useful solution: *bin the outcomes/residuals*.
    - Bin observations with similar fitted values.
    - Take the average of residuals and fitted values.
    - Plot the averages against one another.
- As residual plots in linear regression, we are looking for random pattern around horizontal line.
- **Note**: There is a balance between enough bins to see patterns and enough observations by bins to have stable averages.

Example i

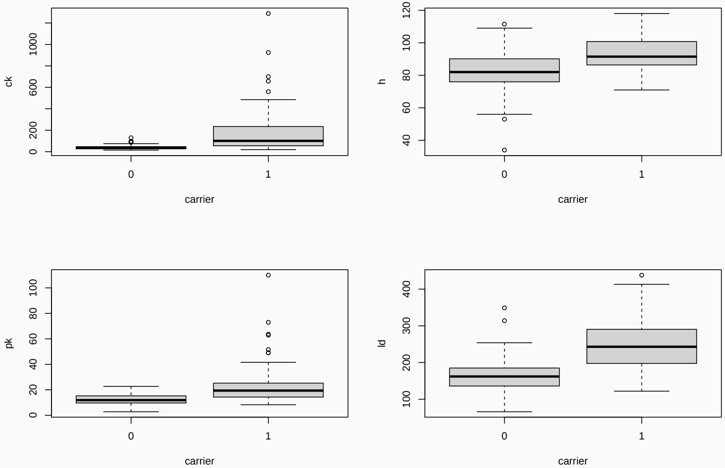
- We will use data on Duchenne Muscular Dystrophy (DMD).
  - Can be downloaded from `http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets`
- Goal of the study was to develop a screening program for female relatives of boys with DMD.
- **Outcome**: Carrier status
- **Covariates**: serum markers: creatine kinase (`ck`), hemopexin (`h`), pyruvate kinase (`pk`) and lactate dehydroginase (`ld`).

Example ii

```r
# Import dataset into R
data_dmd <- read.csv(”DMD.csv”)
# Remove rows with missing values
data_dmd <- na.omit(data_dmd)

# Explore data
par(mfrow = c(2, 2))
boxplot(ck ~ carrier, data = data_dmd)
boxplot(h ~ carrier, data = data_dmd)
boxplot(pk ~ carrier, data = data_dmd)
boxplot(ld ~ carrier, data = data_dmd)
```

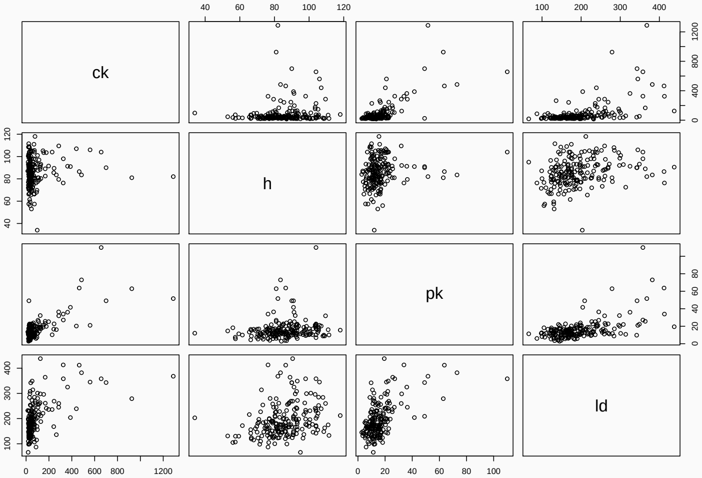
Example iii


ck
0
200
600
1000
carrier
0
1
h
40
60
80
100
120
carrier
0
1
pk
0
20
40
60
80
100
carrier
0
1
ld
100
200
300
400
carrier
0
1

Example iv

```r
# Pairs plot----
# Useful for pairwise comparisons
with(data_dmd, pairs(cbind(ck, h, pk, ld)))
```

Example v


ck
h
pk
ld
40
60
80
100
120
100
200
300
400
1200
800
400
0
120
100
80
60
40
80
60
40
20
0
400
300
200
100
0
200
400
600
800
1200
0
20
40
60
80
100

Example vi

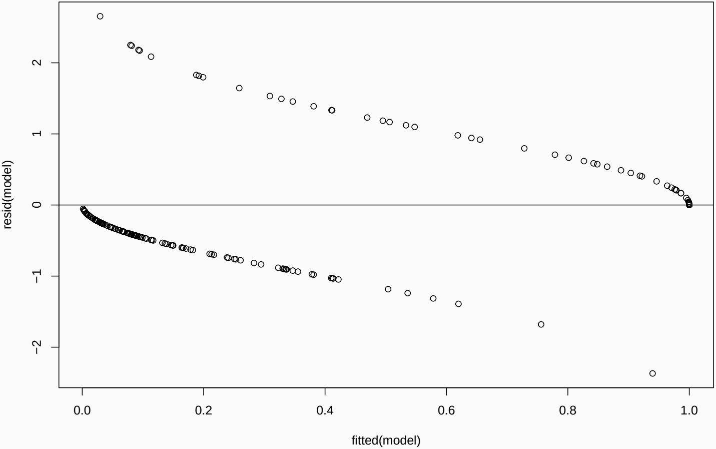
```r
model <- glm(carrier ~ ck + h, data = data_dmd,
             family = "binomial")
confint(model)
```

```
##                    2.5 %       97.5 %
## (Intercept) -20.76823776 -10.43024757
## ck            0.04058575   0.08519017
## h             0.07813791   0.17837069
```

Example vii

```r
# Plot residuals and probabilities (no binning)
plot(fitted(model), resid(model))
abline(h = 0)
```

Example viii



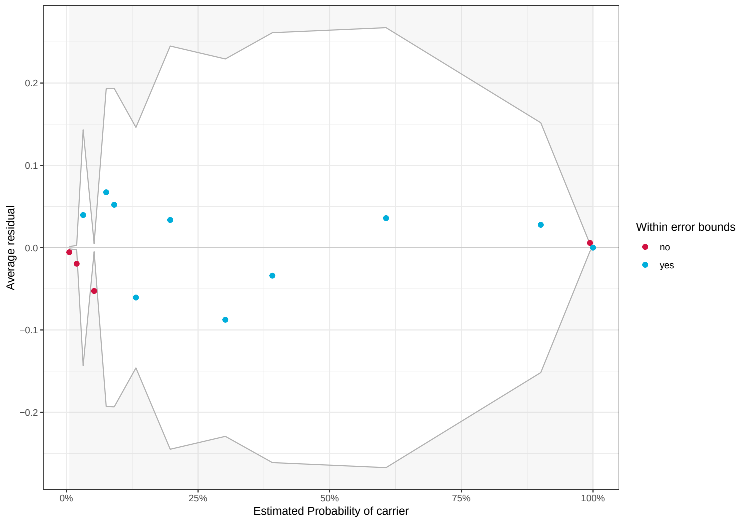
Example ix

```r
# We will use the 'performance' package
library(performance)

# By default: residuals vs fitted probs
#             sqrt(n) bins (~14 bins)
binned_residuals(model)
```

```
## Warning: Probably bad model fit. Only about
71% of the residuals are inside the error bounds.
```

Example x

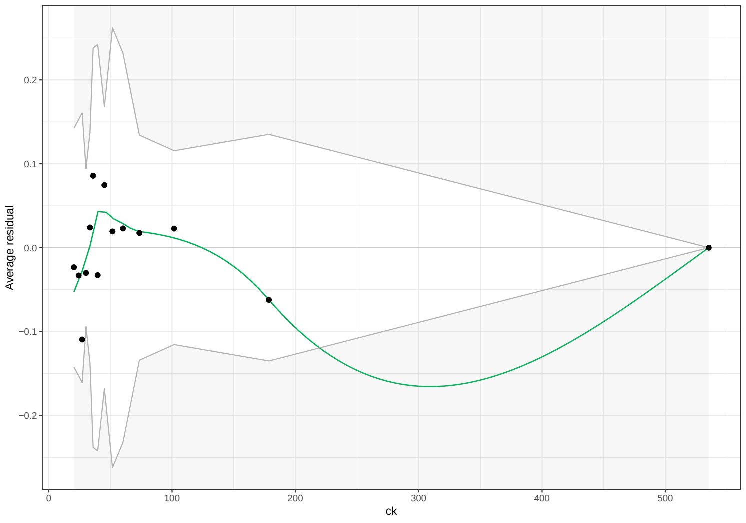

0.2
0.1
0.0
−0.1
−0.2
Average residual
0%
25%
50%
75%
100%
Estimated Probability of carrier
Within error bounds
● no
● yes

Example xi

```r
# Use 'term' to plot against covariate
binned_residuals(model, term = ”ck”)
```

```
## Ok: About 100% of the residuals are inside the
error bounds.
```

Example xii

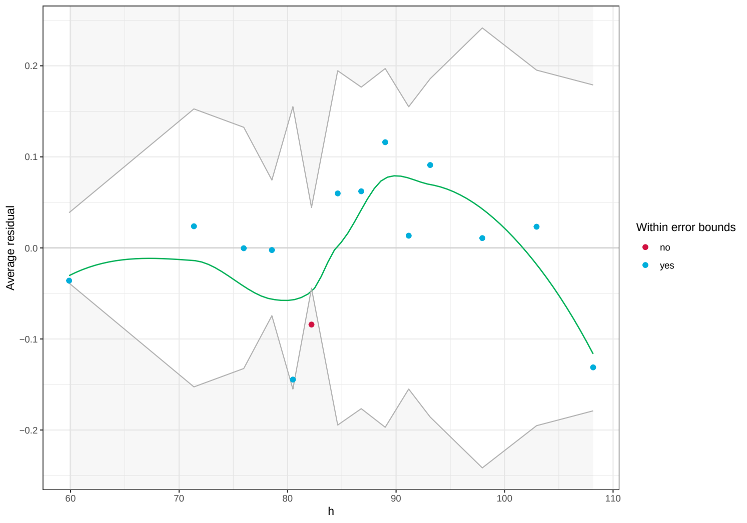


Example xiii

```
binned_residuals(model, term = ”h”)
```

```
## Warning: About 93% of the residuals are inside
the error bounds (~95% or higher would be good).
```

Example xiv



Average residual
0.2
0.1
0.0
−0.1
−0.2
60
70
80
90
100
110
h
Within error bounds
● no
● yes

Example xv

- We have evidence of poor model fit (from binned residuals vs fitted probabilities).
    - But the evidence is weak.
- It may be driven by non-linearity of the effect of h on the log-odds.
    - Or it could be driven by a missing covariate.

Other considerations i

- **Calibration**: Are the estimated probabilities close to empirical probabilities?
    - Hosmer-Lemeshow, Brier score
- **Discrimination**: Are cases more likely to be given large scores (or large probabilities) than non-cases?
    - Area under the ROC curve (AUC), Percentage of Correct Predictions (PCP)
    - **Note**: the AUC is not a very sensitive measure of model performance.

Other considerations ii

```r
performance_hosmer(model)
```

```
## # Hosmer-Lemeshow Goodness-of-Fit Test
##
##   Chi-squared: 3.305
##            df: 8
##       p-value: 0.914
```

```r
# Quadratic score = Brier score
performance_score(model)
```

Other considerations iii

```
## # Proper Scoring Rules
##
## logarithmic:    -Inf
##   quadratic: 8.1783
##   spherical: 0.0280
```

```r
performance_roc(model)
```

```
## AUC: 92.73%
```

```r
performance_pcp(model)
```

Other considerations iv

```
## # Percentage of Correct Predictions from
Logistic Regression Model
##
## Full model: 81.53% [76.07% - 86.99%]
## Null model: 54.78% [47.78% - 61.79%]
##
## # Likelihood-Ratio-Test
##
## Chi-squared: 133.685
## p-value: 0.000
```

References

- Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories.* Cambridge University Press.
- Harrell Jr, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis.* Springer.
- Harrell Jr, F. E. (2019). *Biostatistics for Biomedical Research.* Course notes available online: `http://hbiostat.org/bbr/`.