Regression

BRSM

Simple Linear Regression

- Interval/ratio scale predictors and outcome variables

Scatterplot

Imagine a line through these points that capture the correlation you're thinking about

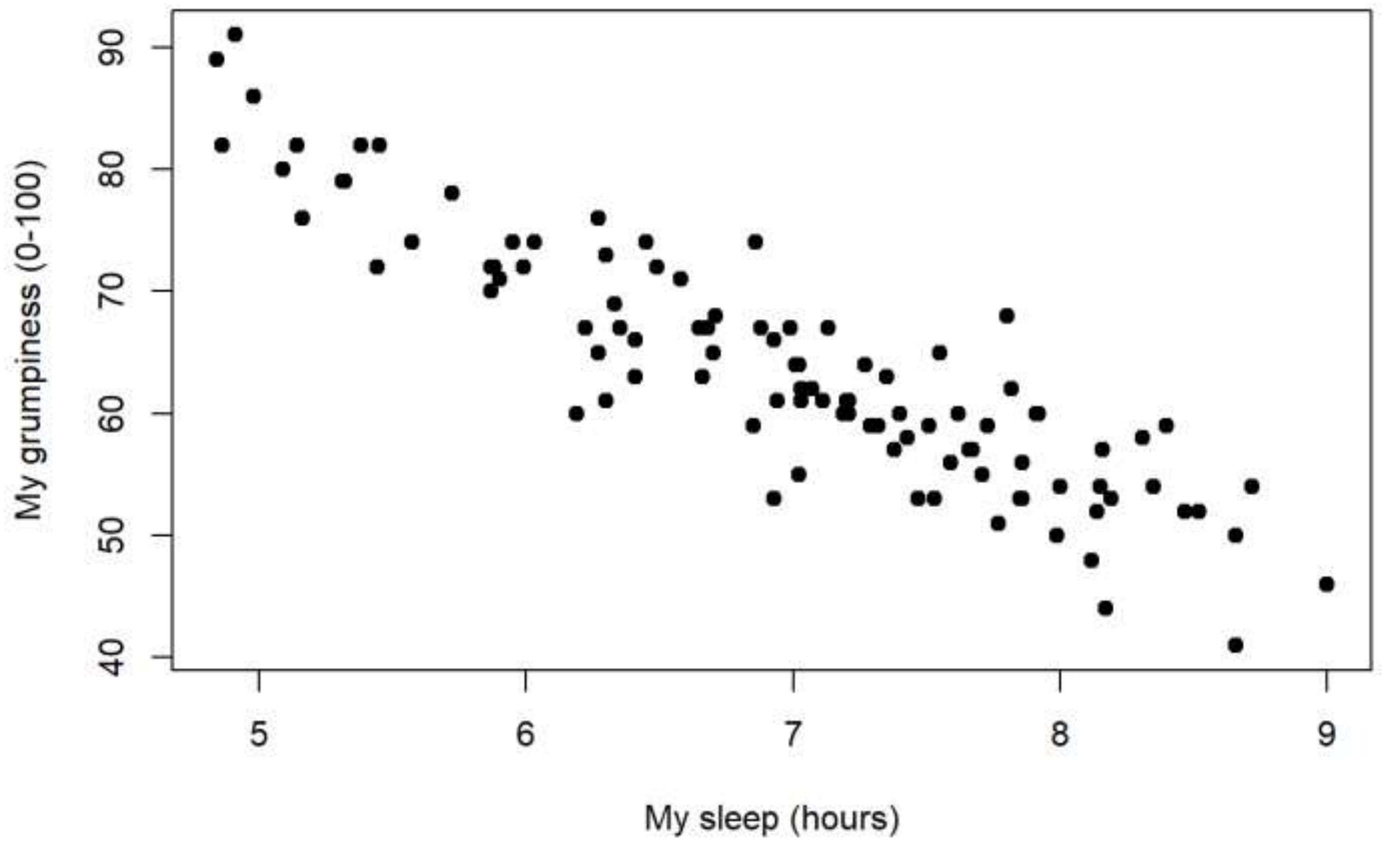
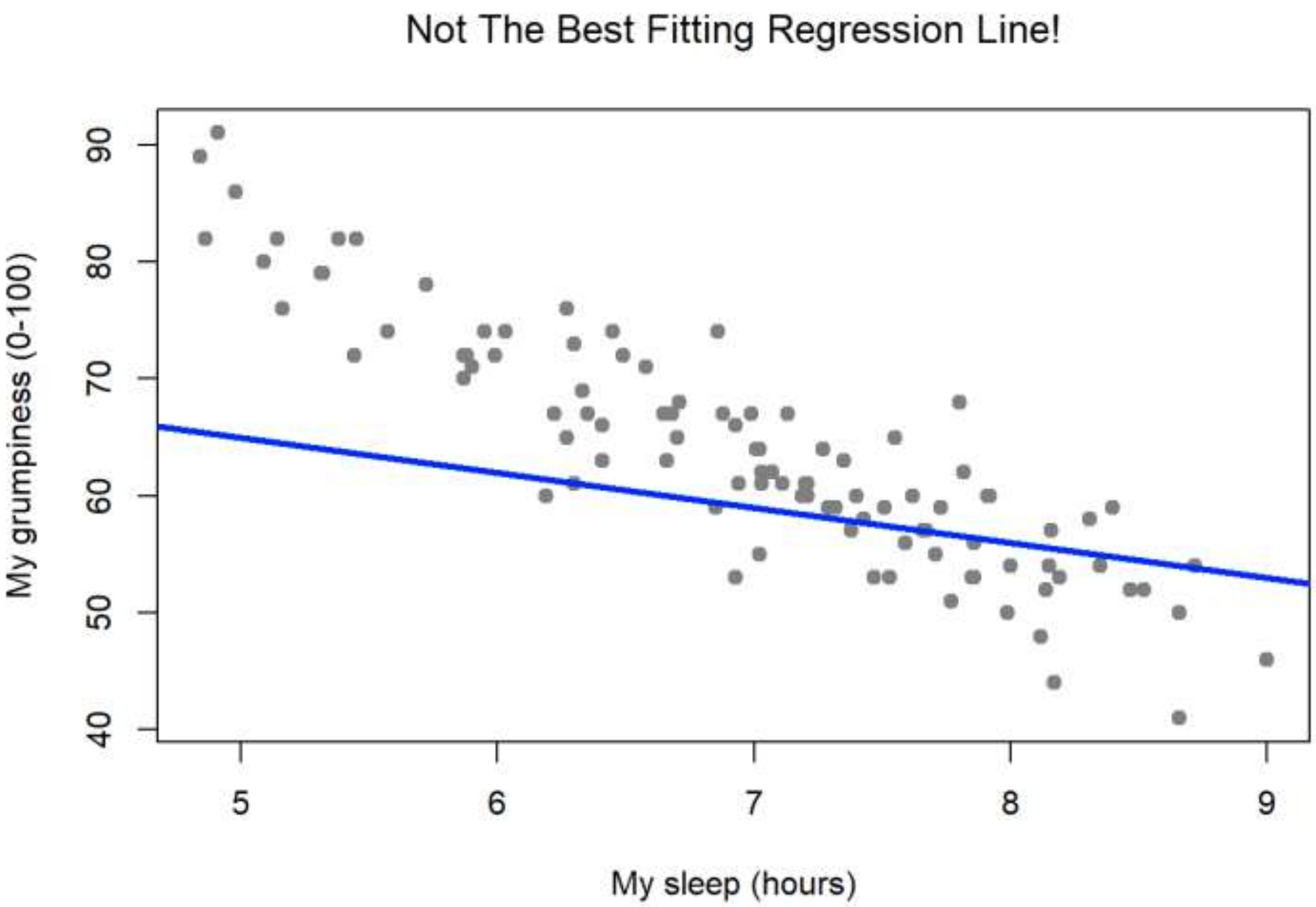


Figure 15.1: Scatterplot showing grumpiness as a function of hours slept.

Best-fitting Regression line

A poor-fitting line



Simple linear regression

- Related to the idea of correlations

$$y = mx + c$$

$$\hat{Y}_i = b_1 X_i + b_0$$

$$\epsilon_i = Y_i - \hat{Y}_i$$

$$Y_i = b_1 X_i + b_0 + \epsilon_i$$

Hat --> predicted
b1 --> regression coefficient
Error -->

Residuals related to the best-fitting regression line

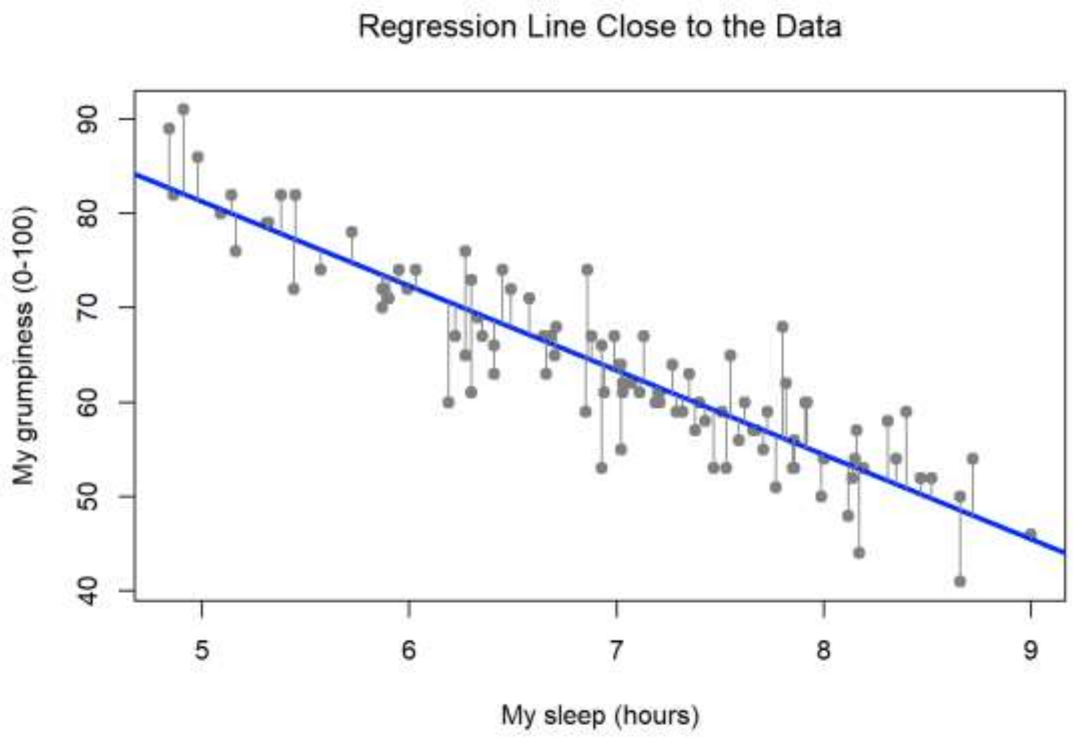
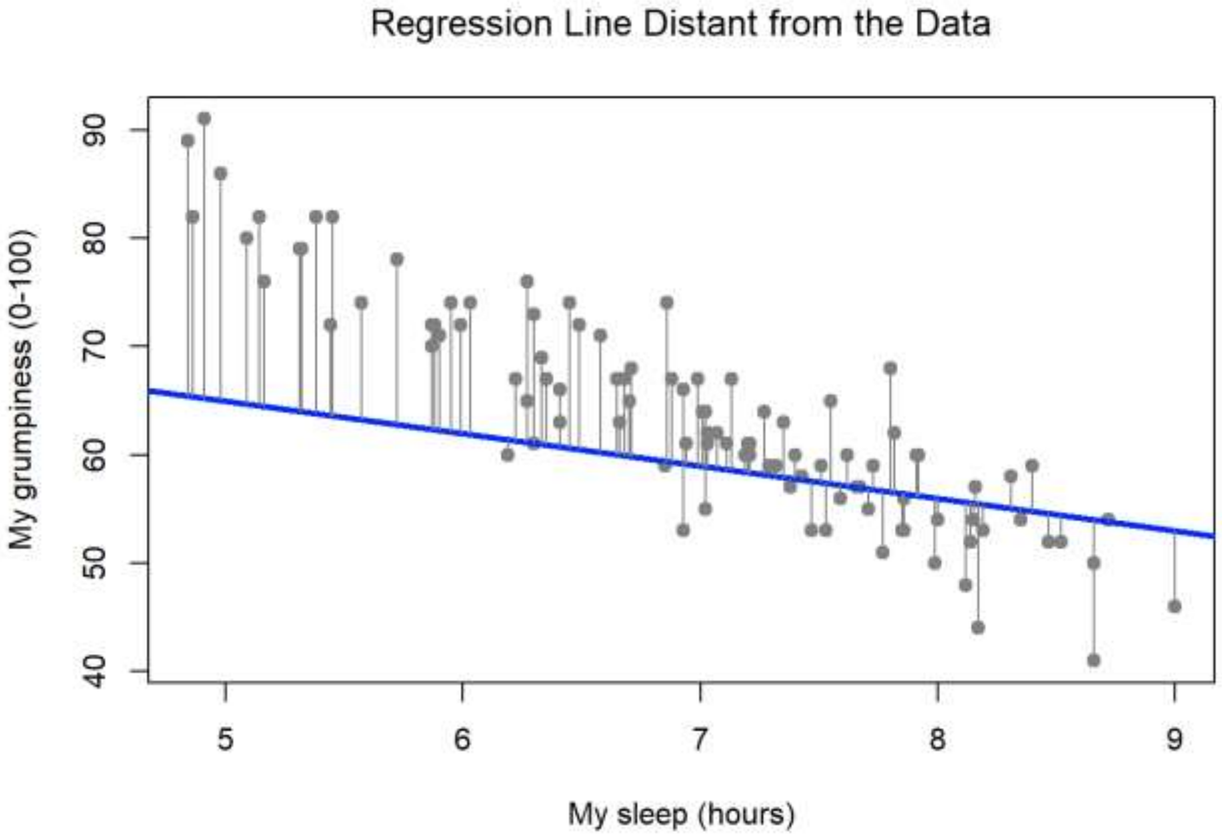


Figure 15.4: A depiction of the residuals associated with the best fitting regression line

Residuals related to a poor-fitting line



How do we estimate the regression coefficients?

- Intuition regarding residuals?
- Small residuals
- Quantity to minimize: sum of squares of errors (residuals)
- This is called Ordinary Least Squares Regression
- Many other ways to estimate regression coefficients

R formula

```r
regression.1 <- lm( formula = dan.grump ~ dan.sleep,
                    data = parenthood )
```

```r
print( regression.1 )
```

```
##
## Call:
## lm(formula = dan.grump ~ dan.sleep, data = parenthood)
##
## Coefficients:
## (Intercept)    dan.sleep
##     125.956       -8.937
```

$$\hat{Y}_i = -8.94\ X_i + 125.96$$

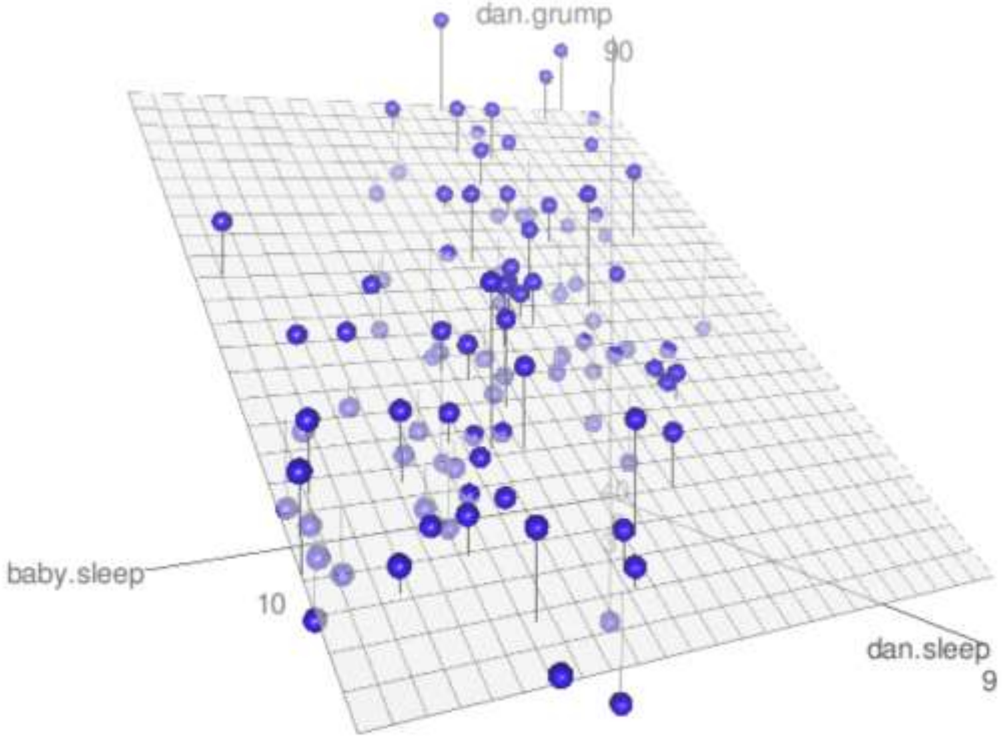
Play time: guess the regression

- https://sophieehill.shinyapps.io/eyeball-regression/

Multiple linear regression (MLR)

- When you have more than one pr

$$Y_i = b_2 X_{i2} + b_1 X_{i1} + b_0 + \epsilon_i$$



R formula

```r
regression.2 <- lm( formula = dan.grump ~ dan.sleep + baby.sleep,
                    data = parenthood )
```

```r
print( regression.2 )
```

```
##
## Call:
## lm(formula = dan.grump ~ dan.sleep + baby.sleep, data = parenthood)
##
## Coefficients:
## (Intercept)    dan.sleep   baby.sleep
##   125.96557     -8.95025      0.01052
```

MLR with k variables

$$Y_i = \left(\sum_{k=1}^{K} b_k X_{ik} \right) + b_0 + \epsilon_i$$

How do you know if the regression does a good job?

```
##
## Call:
## lm(formula = dan.grump ~ dan.sleep + baby.sleep, data = parenthood)
##
## Coefficients:
## (Intercept)    dan.sleep   baby.sleep
##   125.96557     -8.95025      0.01052
```

- Can you infer how good the regression line fit is based on the coefficients?
- No, these just help you predict Y, how good this prediction is needs to be quantified.

R squared

$$\mathrm{SS}_{res} = \sum_i (Y_i - \hat{Y}_i)^2$$

Sum of squared residuals (SSR), which we hope is small.

How small should this be? What do we compare against?

$$\mathrm{SS}_{tot} = \sum_i (Y_i - \bar{Y})^2$$

The outcome variable Y itself is quite variable. If the SSR << the variability in Y, that is a good sign. If the SSR is the same as the

R squared

$$SS_{res} = \sum_i (Y_i - \hat{Y}_i)^2$$

$$SS_{tot} = \sum_i (Y_i - \bar{Y})^2$$

So construct something that is 0 if the fit is the worst and 1 if the fit is the best.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

The coefficient of determination

The proportion of variance in the outcome variable accounted for by the

the relationship between regression and correlati

- The squared Pearson correlation and the R square value from the regression are the same for the case of one predictor.

What is one easy way to improve R square?

- Add more predictors!
- The R square will never decrease by adding more predictors.
- However, this added complexity of the model should be accounted for in your measure of goodness of fit.
- Adjusted R square: constructed such that additional variables will improve adj R square only if the added variables significant ... you'd expect by

$$\text{adj. } R^2 = 1 - \left(\frac{SS_{res}}{SS_{tot}} \times \frac{N-1}{N-K-1} \right)$$

What should you report: R square or Adj. R square?

- R square: straightforward to interpret as the proportion of variance in the outcome variable accounted for by the predictors but does not account for complexity and added degrees of freedom due to added predictor variables.
- Adj . R square: not straightforward to interpret but is a measure of goodness of fit that is not biased by added complexity of the model.

Next: hypothesis tests for regression models and coefficients

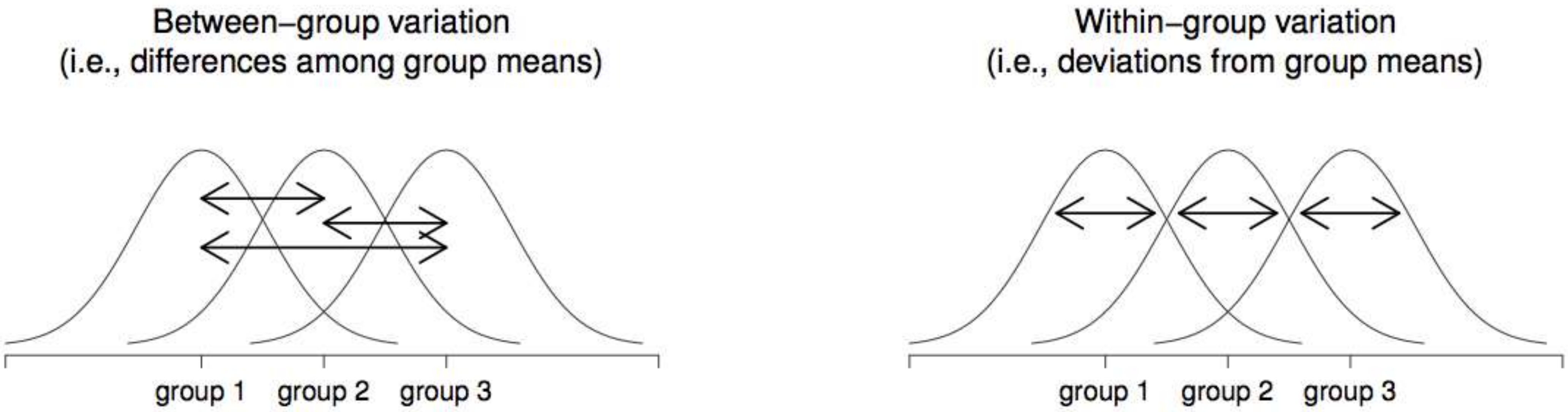
- So far: interpreting regression coefficients, and evaluating overall goodness of fit, but we do not know if a regression coefficient of 3.4 for instance is statistically significant (i.e., statistically meaningfully greater than 0).
- We also need to do a statistical test for the model as a whole by comparing it against a trivial model, as it is possible that the use of a more trivial model can also lead to comparable R squares in some situations.

Hypothesis tests for the entire regression model

- The null mod $H_0 : Y_i = b_0 + \epsilon_i$
- The alternat $H_1 : Y_i = \left(\sum_{k=1}^{K} b_k X_{ik} \right) + b_0 + \epsilon_i$
- To construct the test, we start by dividing the total sum of squares of the outcome variable just as it is done in ANOVA

$$\mathrm{SS}_{mod} = \mathrm{SS}_{tot} - \mathrm{SS}_{res}$$

A reminder about ANOVA



A reminder about ANOVA

$$\mathrm{SS}_{tot} = \sum_{k=1}^{G} \sum_{i=1}^{N_k} \left(Y_{ik} - \bar{Y}\right)^2$$

$$\mathrm{SS}_w = \sum_{k=1}^{G} \sum_{i=1}^{N_k} \left(Y_{ik} - \bar{Y}_k\right)^2$$

$$\begin{aligned} \mathrm{SS}_b &= \sum_{k=1}^{G} \sum_{i=1}^{N_k} \left(\bar{Y}_k - \bar{Y}\right)^2 \\ &= \sum_{k=1}^{G} N_k \left(\bar{Y}_k - \bar{Y}\right)^2 \end{aligned}$$

$$\mathrm{SS}_w + \mathrm{SS}_b = \mathrm{SS}_{tot}$$

A reminder about ANOVA

	df	sum of squares	mean squares	F-statistic	p-value
between groups	$\mathrm{df}_b = G - 1$	$\mathrm{SS}_b = \sum_{k=1}^{G} N_k (\bar{Y}_k - \bar{Y})^2$	$\mathrm{MS}_b = \frac{\mathrm{SS}_b}{\mathrm{df}_b}$	$F = \frac{\mathrm{MS}_b}{\mathrm{MS}_w}$	[complicated]
within groups	$\mathrm{df}_w = N - G$	$\mathrm{SS}_w = \sum_{k=1}^{G} \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y}_k)^2$	$\mathrm{MS}_w = \frac{\mathrm{SS}_w}{\mathrm{df}_w}$	-	-

Large F --> ??

Back to linear regression and sum of squares

$$\mathrm{SS}_{mod} = \mathrm{SS}_{tot} - \mathrm{SS}_{res}$$

$$df_{mod} = K.$$

$$df_{res} = N - K - 1.$$

$$F = \frac{\mathrm{MS}_{mod}}{\mathrm{MS}_{res}}$$

Similar interpretation as in the ANOVA case. High value of F --> the

So we have just tested the regression model as a whole

- If the F test is not significant, then either the model is a poor one or your data has issues.
- If it is significant, it still doesn't mean you know for sure your predictors all explain the outcome. Need to do statistical tests for individu

```
> print( regression.2 )

Call:
lm(formula = dan.grump ~ dan.sleep + baby.sleep, data = parenthood)

Coefficients:
(Intercept)    dan.sleep   baby.sleep
  125.96557     -8.95025      0.01052
```

Testing for individual regression coefficients

- CLT
- Normally distributed sampling distribution of the estimator of b, centered on b.
- If we can then come up with a standard error for this estimator, then we can construct a t-statistic
- Turns out we can do this, a complicated formula, but note that this

$$
\begin{aligned}
H_0 &: \quad b = 0 \\
H_1 &: \quad b \neq 0
\end{aligned}
\qquad\qquad t = \frac{\hat{b}}{\mathrm{SE}(\hat{b})}
$$

Hypothesis test for coefficients in R

```
> summary( regression.2 )

Call:
lm(formula = dan.grump ~ dan.sleep + baby.sleep, data = parenthood)

Residuals:
     Min       1Q   Median       3Q      Max
-11.0345  -2.2198  -0.4016   2.6775  11.7496
```

Regression assumption: residuals are normally distributed around 0.
So check if median around 0, 1Q and 3Q approx equidistant from 0...

Hypothesis test for coefficients in R

```
> summary( regression.2 )
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 125.96557    3.04095  41.423   <2e-16 ***
dan.sleep    -8.95025    0.55346 -16.172   <2e-16 ***
baby.sleep    0.01052    0.27106   0.039    0.969
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 4.354 on 97 degrees of freedom
```

Hypothesis test for coefficients in R

```
> summary( regression.2 )
```

```
Residual standard error: 4.354 on 97 degrees of freedom
Multiple R-squared: 0.8161,      Adjusted R-squared: 0.8123
F-statistic: 215.2 on 2 and 97 DF,  p-value: < 2.2e-16
```

A global assessment of
the model

Confidence intervals for regression coefficients

$$\mathrm{CI}(b) = \hat{b} \pm \left(t_{crit} \times \mathrm{SE}(\hat{b})\right)$$

```
> confint( object = regression.2,
+          level = .99
+ )
                  0.5 %      99.5 %
(Intercept) 117.9755724 133.9555593
dan.sleep   -10.4044419  -7.4960575
baby.sleep   -0.7016868   0.7227357
```

N-K-1 degrees of freedom
Critical t value for 97.5th percentile, to construct a 95% CI.

How do you compare regression coefficients for predictors that have different units and have totally different scales?

- e.g. Predicting intelligence scores using years of education and income.
- Income: may vary from tens of thousands p.a to several lakhs.
- Years of education: 0-15 years
- Comparing regression coefficients from these predictors would be difficult. Say 0.25 for income and 0.89 for years of education.

$$\beta_X = b_X \times \frac{\sigma_X}{\sigma_Y}$$

- Here, we can use standardized regress[...] score your variables and do the regression)

Interpreting standardized regression coefficients

- IQ ~ b1*income + b2*years of education + b0
- Standardized coefs are usually denotes by betas.
- A change in income by 1 s.d. (of income) corresponds to beta1 s.d change in IQ when years of education is held constant.
- Can directly compare b1 and b2 in terms of how much each variable affects IQ (in terms of IQ s.d.)
- However, 1 s.d. change in income and 1 s.d. change in years of education – are they comparable quantities? This is not too straightforward. So while standardized regression is supposed to help you put different predictors on the same scale, you have

Final section of the basics: The assumptions of linear regression

- Normality: residuals are normally distributed. The variables can be non-normal!
- Linearity: the relationship between X and Y is more or less linear
- Homogeneity of variance: We assume that the residuals are i.i.d with mean 0 and the same s.d. Not easy to test this, but we will check whether the s.d. of the residuals are the same at each level of X and Y instead --> homogeneity of variance.

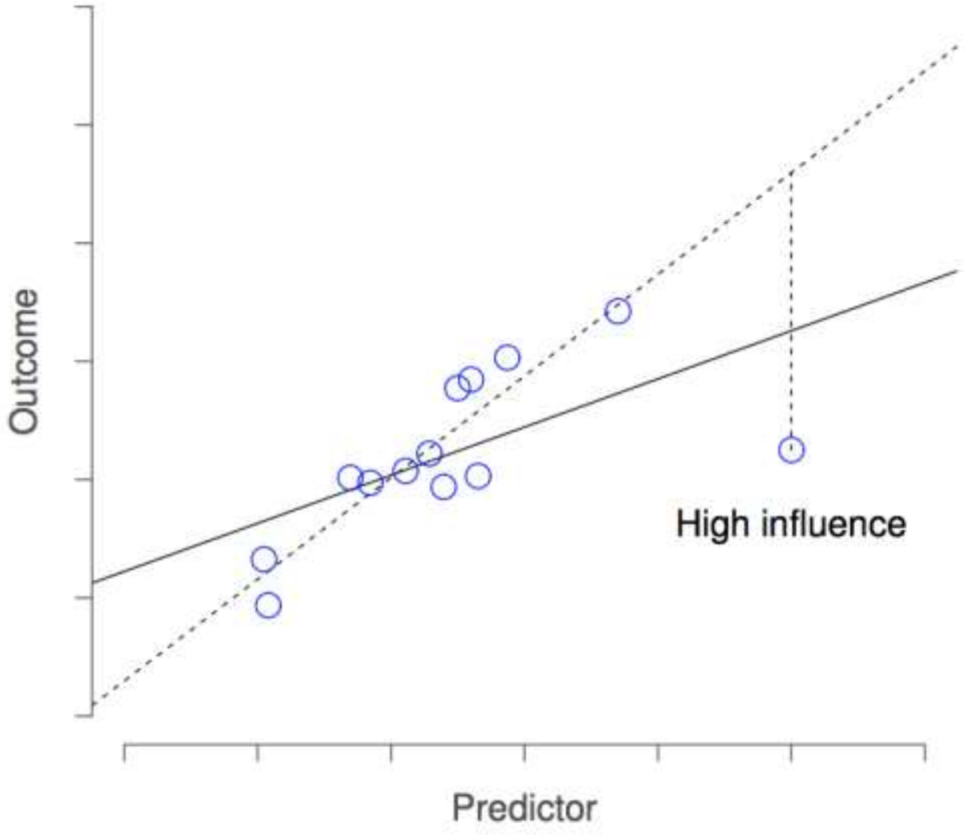
Other desirable features for regression (but not strict assumptions)

- Uncorrelated predictors – collinear/correlated predictors makes it hard to interpret the regression output in many cases.
- No large outliers - Is the regression being influenced heavily by one or two points?

Regression diagnostics

Checking for outlier influence: Cook's distance

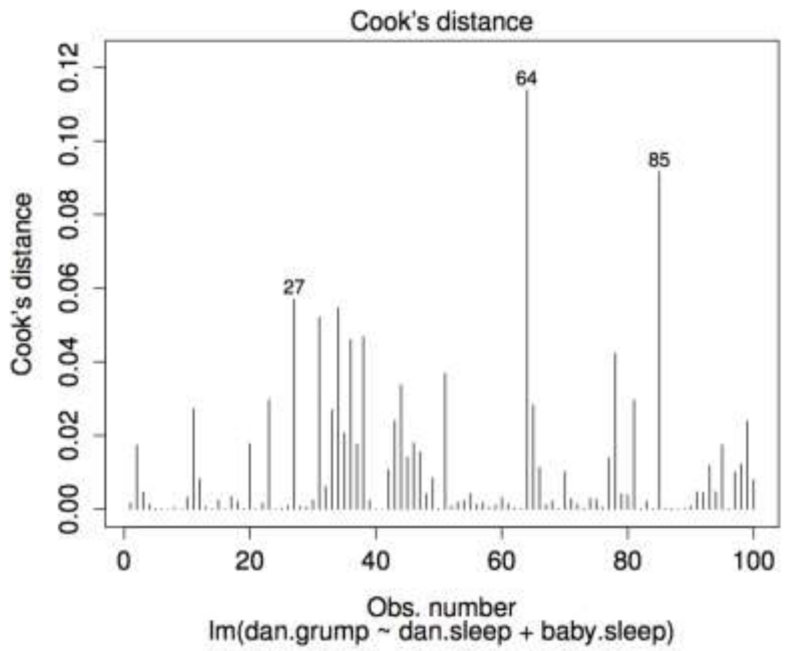
$$D_i = \frac{\epsilon_i^{*2}}{K+1} \times \frac{h_i}{1-h_i}$$



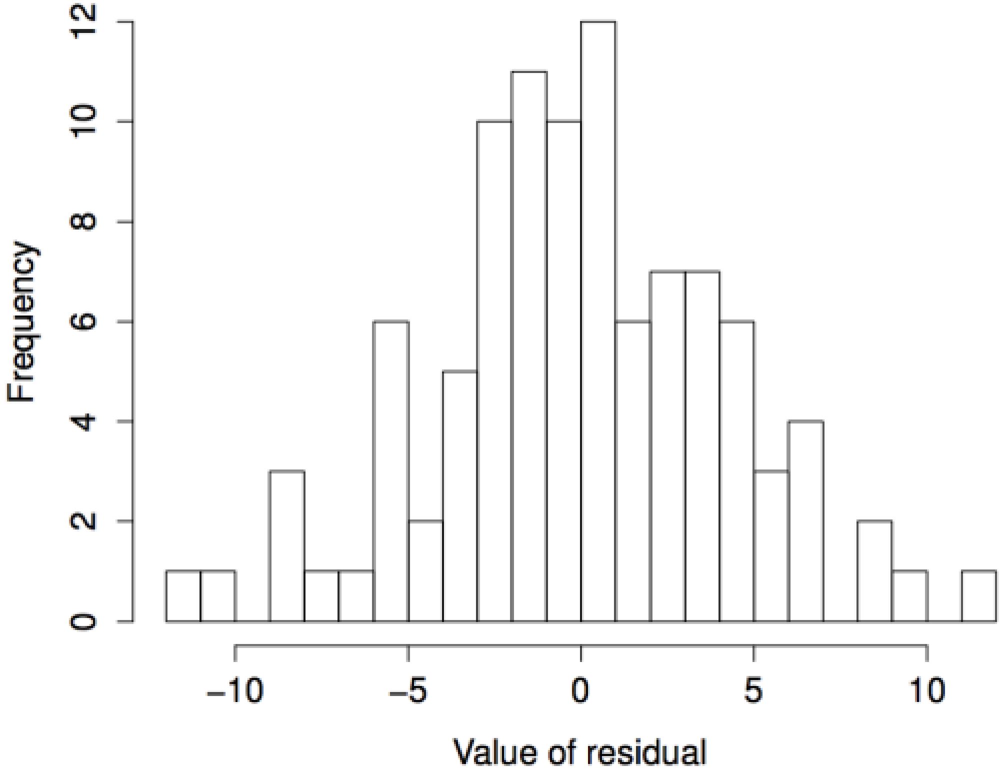
Cook's distance plots: checking for outlier influence

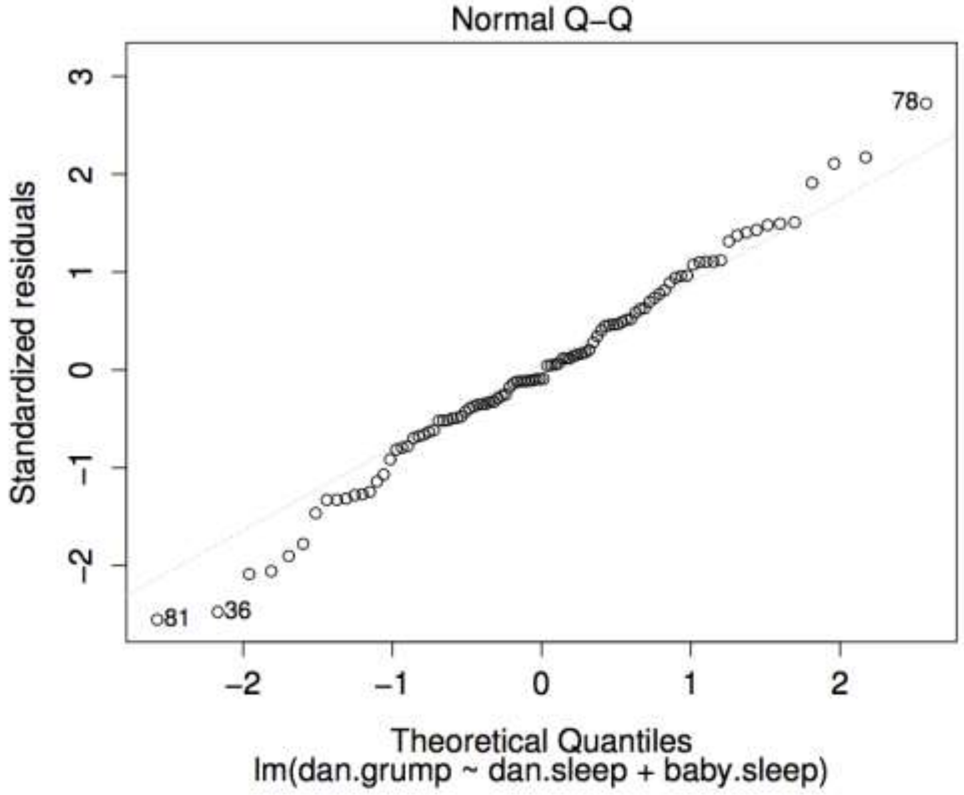
```
> plot(x = regression.2, which = 4)
```

Cook's distance > 1 might indicate problems If you get a point like that, try removing that data point and re-running your regression. If the coefficients and results change by a lot, you can tell that the outlier had a huge influence.



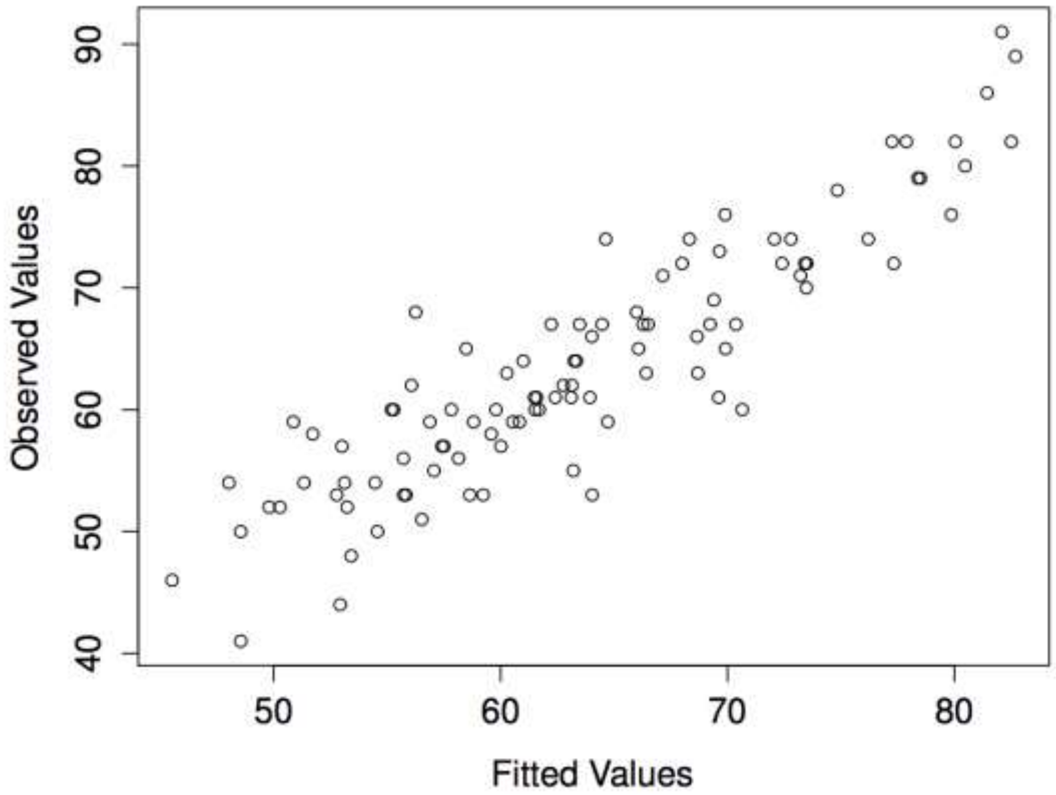
Normality of residuals



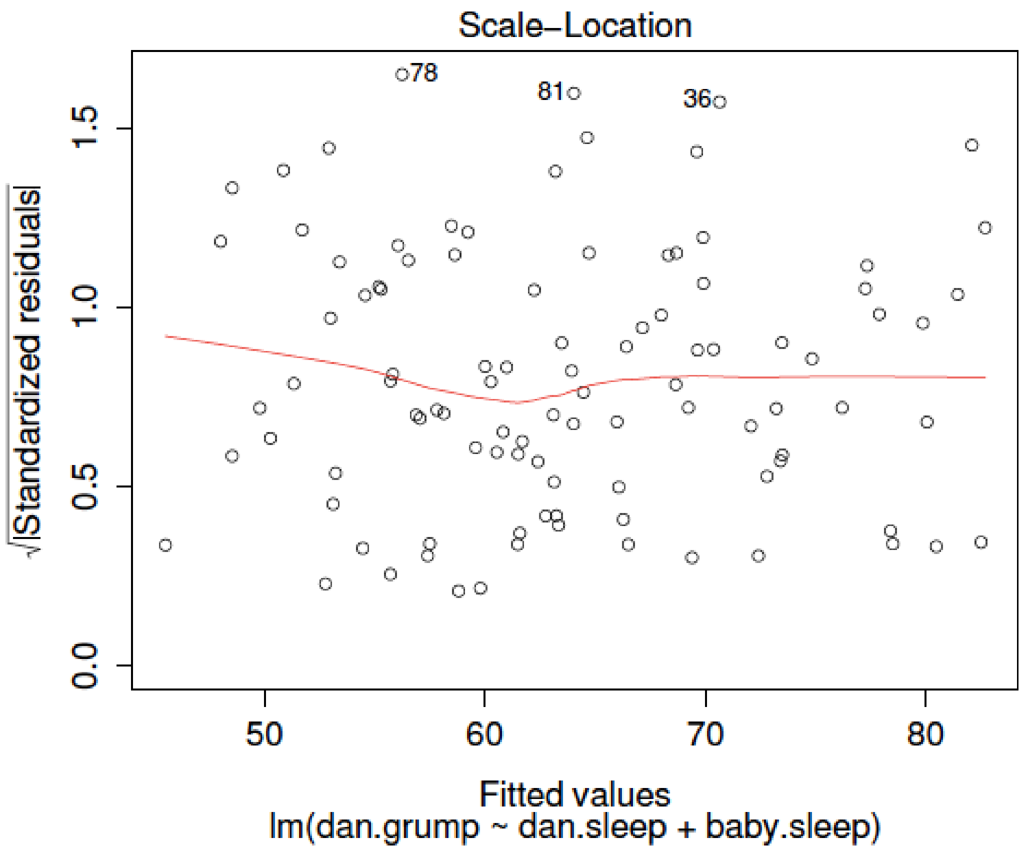


Also do the Shapiro-Wilk test, etc

Checking linearity



Checking for homogeneity of variance




```
> ncvTest( regression.2 )
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.09317511    Df = 1     p = 0.7601788
```

How to deal with violations of the homogeneity of variance assumption?

- The problem? Our SE estimates of the estimators of the regression coefficients will no longer be correct as they are based on this homogeneity assumption.
- So will have to use other estimators for computing this SE
- These have been figured out: using heteroscedasticity corrected covariance matrix
- "sandwich estim

```
> coeftest( regression.2, vcov= hccm )
```

Checking for collinearity

- VIF
- Typical rules of thumb: >5 or 10

$$\mathrm{VIF}_k = \frac{1}{1 - R^2_{(-k)}}$$

```
> regression.3 <- lm( day ~ baby.sleep + dan.sleep + dan.grump, parenthood )
```

```
> vif( regression.3 )
baby.sleep  dan.sleep  dan.grump
  1.651064   6.102337   5.437903
```

Comparing regression models: Model selection and occam's razor

$$\mathrm{AIC} = \frac{\mathrm{SS}_{res}}{\hat{\sigma}^2} + 2K$$

Step regression

- Backward: specify the full model first and then remove predictors one
  with the lowes

```
> full.model <- lm( formula = dan.grump ~ dan.sleep + baby.sleep + day,
+                   data = parenthood
+ )
```

```
> step( object = full.model,        # start at the full model
+       direction = "backward"      # allow it remove predictors but not add them
+ )
```

```
Start:  AIC=299.08
dan.grump ~ dan.sleep + baby.sleep + day
```

Step regression

	Df	Sum of Sq	RSS	AIC
- baby.sleep	1	0.1	1837.2	297.08
- day	1	1.6	1838.7	297.16
<none>			1837.1	299.08
- dan.sleep	1	4909.0	6746.1	427.15

Step regression

```
Step:  AIC=297.08
dan.grump ~ dan.sleep + day

            Df Sum of Sq    RSS    AIC
- day        1       1.6 1838.7 295.17
<none>                   1837.2 297.08
- dan.sleep  1    8103.0 9940.1 463.92
```

Step regression

```
Step:  AIC=295.17
dan.grump ~ dan.sleep

            Df Sum of Sq    RSS    AIC
<none>                   1838.7 295.17
- dan.sleep  1    8159.9 9998.6 462.50
```

Step regression: final chosen model

```
Call:
lm(formula = dan.grump ~ dan.sleep, data = parenthood)

Coefficients:
(Intercept)    dan.sleep
    125.956       -8.937
```

Forward step regression

- Also possible
- The answers from forward and backward regression need not always be the same! So be careful when using this, always use your intuition about interpretability of the resulting models as well in addition to all these numbers and diagnostics.

Comparing two regression models in general

```
> M0 <- lm( dan.grump ~ dan.sleep + day, parenthood )
> M1 <- lm( dan.grump ~ dan.sleep + day + baby.sleep, parenthood )
```

```
> AIC( M0, M1 )
   df      AIC
M0  4 582.8681
M1  5 584.8646
```

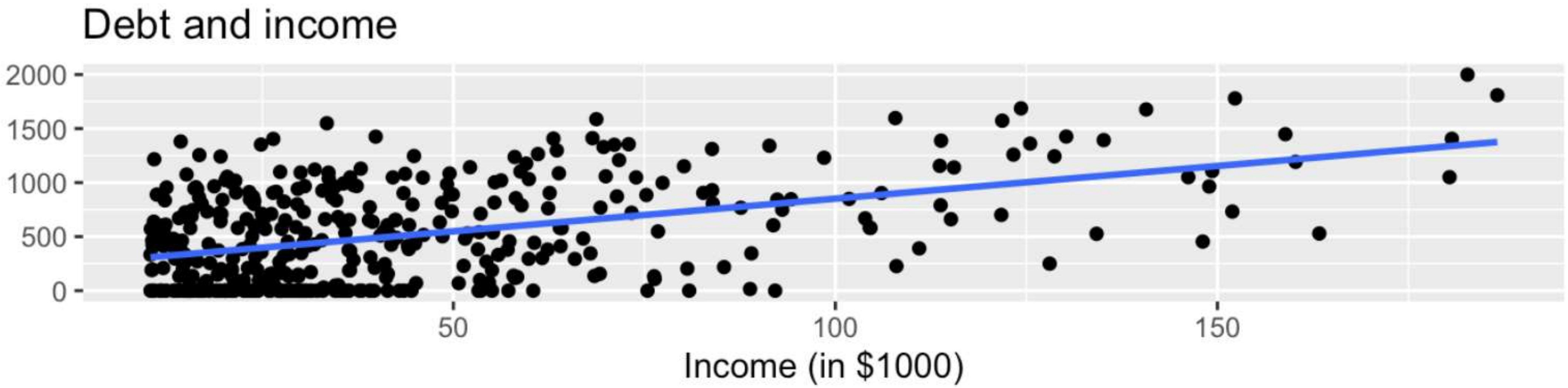
Summary

- Basic ideas in linear regression and how regression models are estimated
- Multiple linear regression
- Measuring the overall performance of a regression model using R2
- Hypothesis tests for regression models
- Calculating confidence intervals for regression coefficients, and standardised coefficients
- The assumptions of regression and how to check them

Other resources

- For a fast-paced and more technical introduction, check out Chapter 1 in Roback and Legler's *Beyond multiple linear regression: Applied generalized linear models and multilevel models in R* (https://github.com/proback/BeyondMLR)
- For an introduction from a Bayesian perspective, Check out Chapters 4 and 5 in McElreath's *Statistical rethinking* (https://osf.io/2h6ut/). You can also find him lecturing on the material in these playlists: https://www.youtube.com/channel/UCNJK6_DZvcMqNSzQdEkzvzA/playlists.

A final note for the day: guess the regression coef for income



Woah! -- Simpson's paradox

TABLE 6.17: Multiple regression results

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-385.179	19.465	-19.8	0	-423.446	-346.912
credit_limit	0.264	0.006	45.0	0	0.253	0.276
income	-7.663	0.385	-19.9	0	-8.420	-6.906

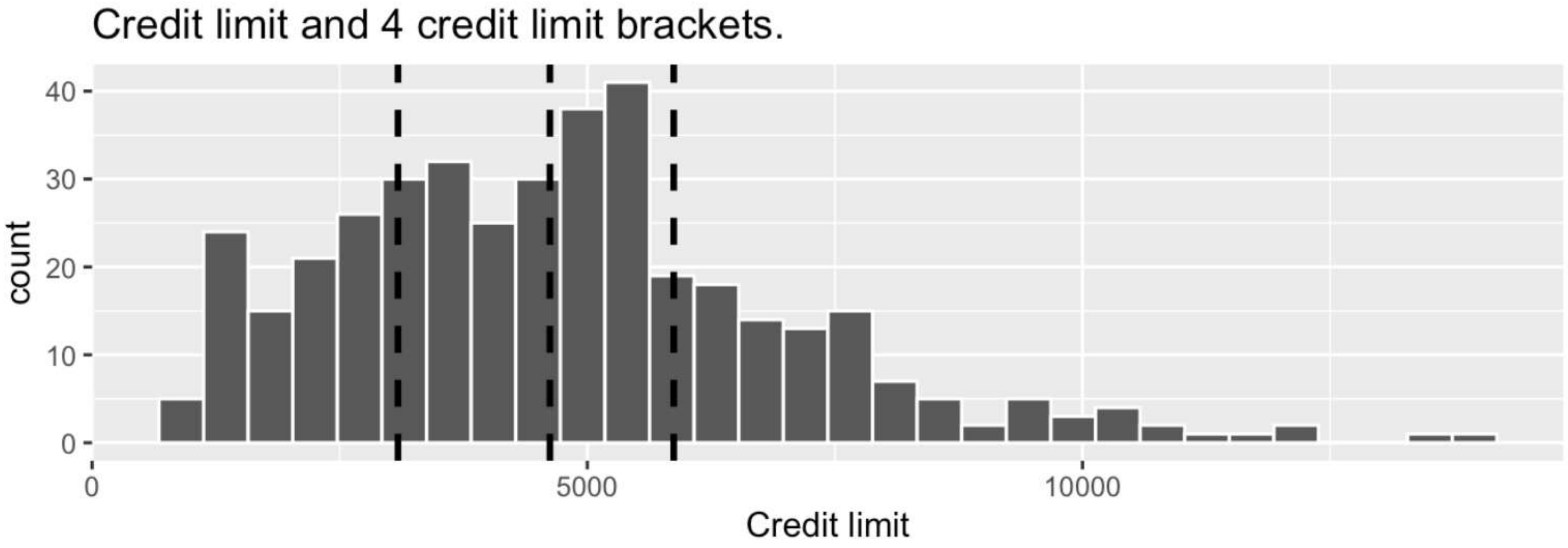
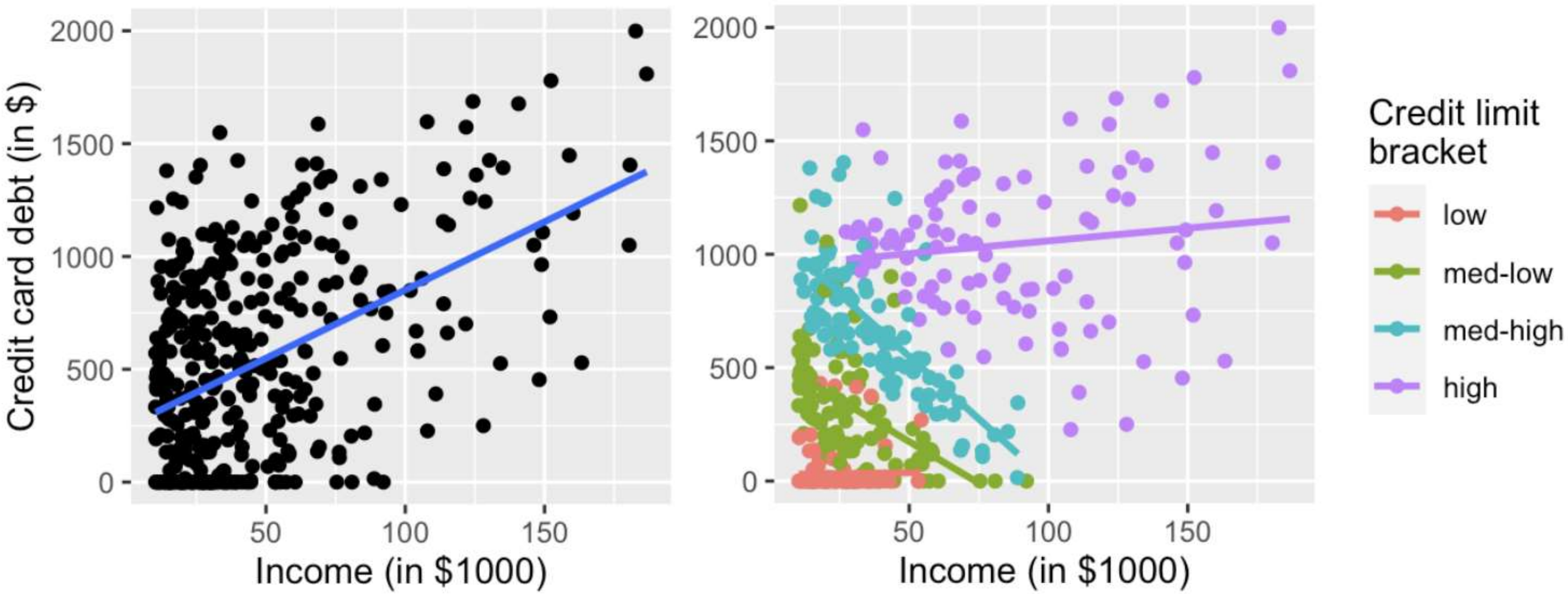


FIGURE 6.10: Histogram of credit limits and brackets.

Two scatterplots of credit car



Next class

- Dealing with other types of variables, interactions, etc
- Practicals
- Now/homework: Simulate some data with y <- b1x1 + b2x2 + b0, add some errors drawn from a normal distribution
- Now fit these simulated data using regression
- Make x1 and x2 correlated, redo, calculate VIF
- Simulate heteroscedasticity? Redo normal regression and compare with regression using the heteroscedasticity corrected covariance matrix option and compare the results.