

Linear models

Example dataset: paper planes flying experiment

```
library(paperplanes)
head(paperplanes)
```

	id	hour	person	gender	age	plane	paper	distance
1	1	[17,18)	Roland	male	30	Standard80	80	7.8
2	2	[17,18)	Astrid	female	30	Concorde120	120	2.7
3	3	[17,18)	Roland	male	30	Standard120	120	9.2
4	4	[17,18)	Isabella	female	48	Standard120	120	6.0
5	5	[17,18)	Fabienne	female	17	Standard120	120	7.3
6	6	[17,18)	Fabienne	female	17	Standard120	120	7.8

Questions

- ▶ What is the relationship between age and distance flown?

Questions

- ▶ What is the relationship between age and distance flown?
- ▶ Do adults achieve longer distances?

Questions

- ▶ What is the relationship between age and distance flown?
- ▶ Do adults achieve longer distances?
- ▶ Can we predict distance flown from participant's age? How well?

Always plot your data first!

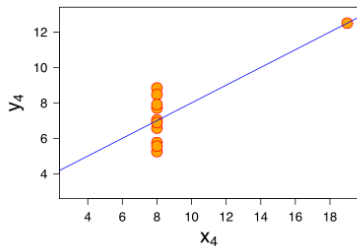
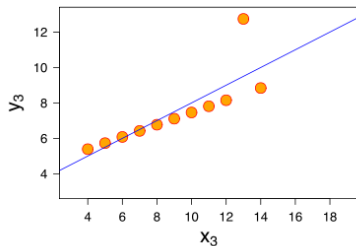
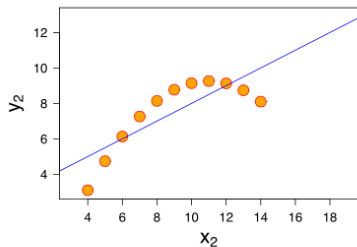
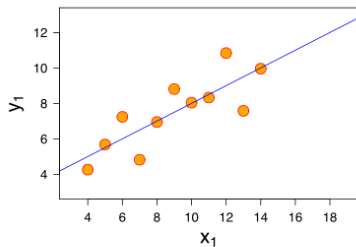
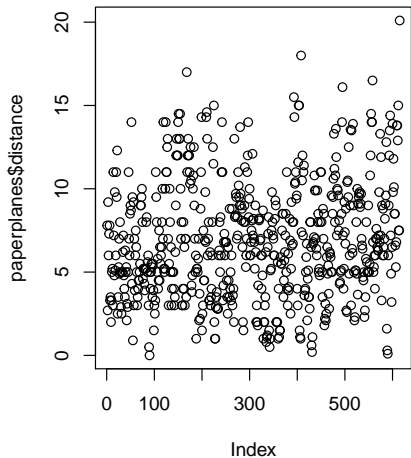


Figure 1:

Exploratory Data Analysis (EDA)

Outliers

```
plot(paperplanes$distance)
```



Outliers impact on regression

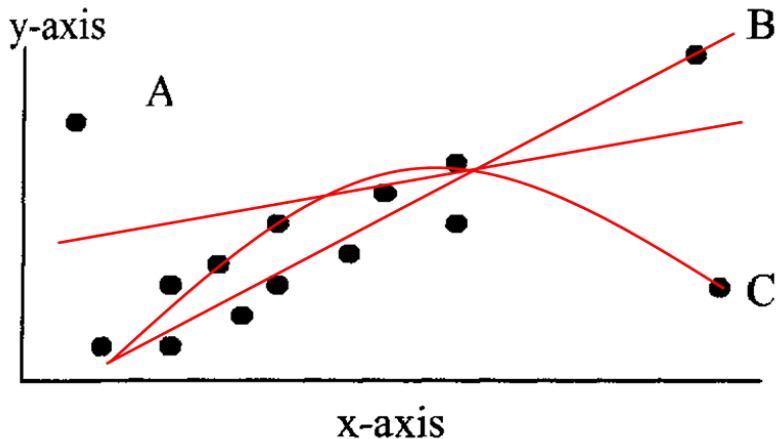
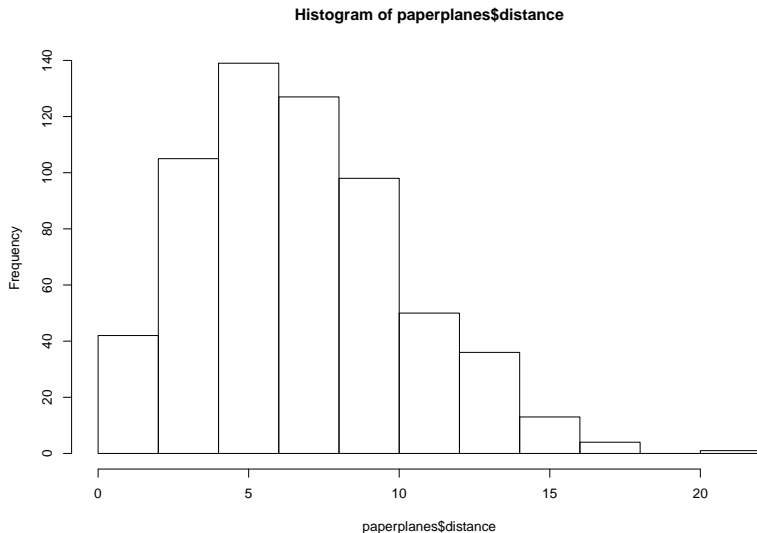


Figure 2:

See <http://rpsychologist.com/d3/correlation/>

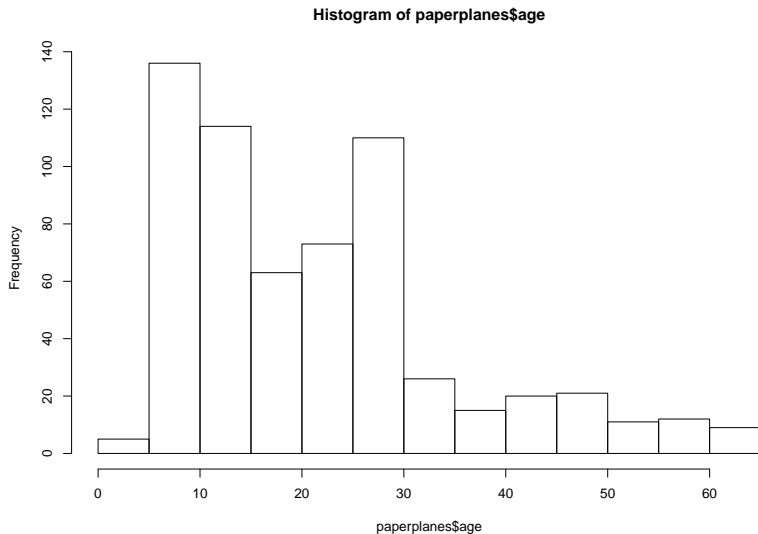
Histogram of response variable

```
hist(paperplanes$distance)
```



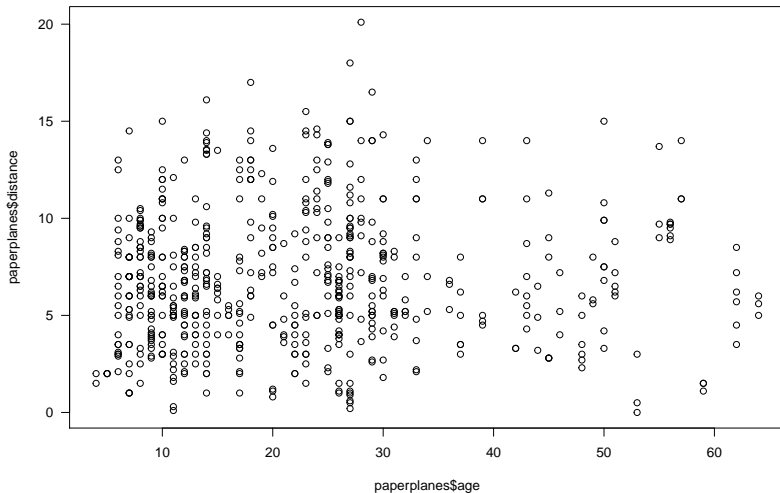
Histogram of predictor variable

```
hist(paperplanes$age)
```



Scatterplot

```
plot(paperplanes$age, paperplanes$distance, las = 1)
```



Now fit model

Hint: `lm`

Now fit model

Hint: `lm`

```
m1 <- lm(distance ~ age, data = paperplanes)
```

What does this mean?

Call:

```
lm(formula = distance ~ age, data = paperplanes)
```

Residuals:

Min	1Q	Median	3Q	Max
-7.1929	-2.6014	-0.3789	2.1572	13.1658

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.64440	0.26982	24.626	<2e-16 ***
age	0.01035	0.01040	0.996	0.32

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.504 on 613 degrees of freedom

Multiple R-squared: 0.001614, Adjusted R-squared: -1.434e-05

F-statistic: 0.9912 on 1 and 613 DF, p-value: 0.3198

Retrieving model coefficients

```
coef(m1)
```

```
(Intercept)          age  
 6.64439782  0.01034968
```

Tidy up model coefficients with broom

```
library(broom)
tidy(m1)
```

	term	estimate	std.error	statistic	p.value
1	(Intercept)	6.64439782	0.26981611	24.6256529	1.286330e-93
2	age	0.01034968	0.01039555	0.9955879	3.198432e-01

```
glance(m1)
```

	r.squared	adj.r.squared	sigma	statistic	p.value	df	l
1	0.001614348	-1.43402e-05	3.503736	0.9911952	0.3198432	2	-164
	AIC	BIC	deviance	df.residual			
1	3291.502	3304.767	7525.289	613			

Confidence intervals

```
confint(m1)
```

	2.5 %	97.5 %
(Intercept)	6.11452177	7.17427388
age	-0.01006553	0.03076489

Using effects package

```
library(effects)  
summary(allEffects(m1))
```

model: distance ~ age

age effect

age

4	20	30	50	60
6.685797	6.851391	6.954888	7.161882	7.265379

Lower 95 Percent Confidence Limits

age

4	20	30	50	60
6.223509	6.570601	6.634085	6.528536	6.443633

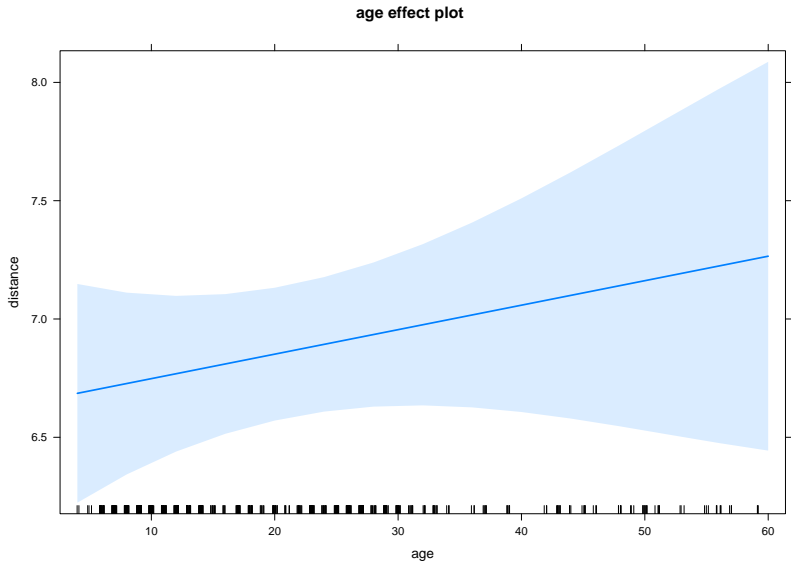
Upper 95 Percent Confidence Limits

age

4	20	30	50	60
7.148084	7.132182	7.275692	7.795228	8.087125

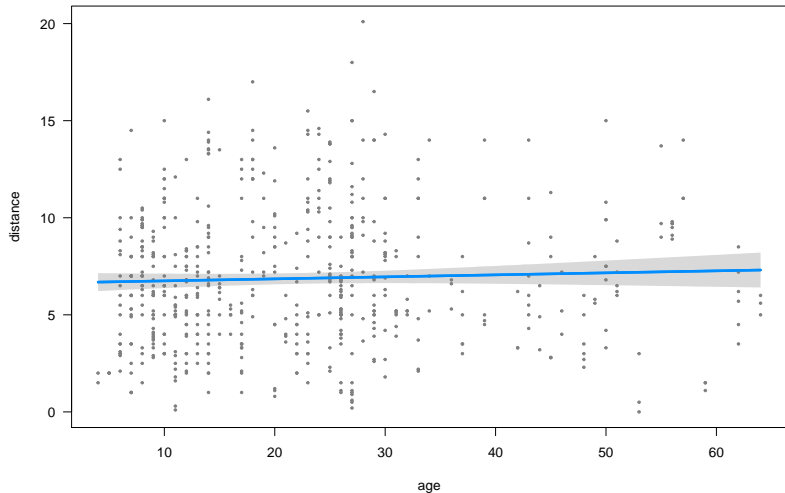
Plot effects

```
plot(allEffects(m1))
```



Plot model (visreg)

```
library(visreg)  
visreg(m1)
```



Linear model assumptions

- ▶ Linearity (transformations, GAM...)

Linear model assumptions

- ▶ Linearity (transformations, GAM. . .)
- ▶ Residuals:

Linear model assumptions

- ▶ Linearity (transformations, GAM...)
- ▶ Residuals:
 - ▶ Independent

Linear model assumptions

- ▶ Linearity (transformations, GAM...)
- ▶ Residuals:
 - ▶ Independent
 - ▶ Equal variance

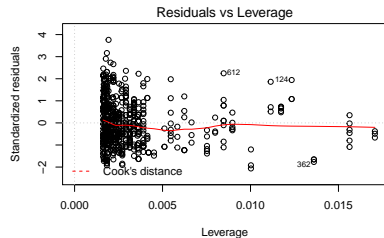
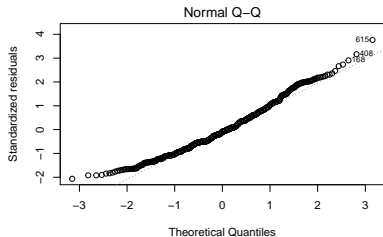
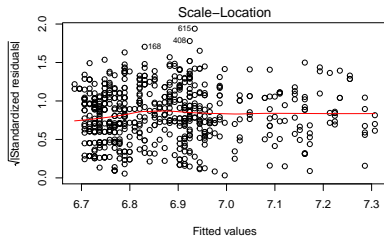
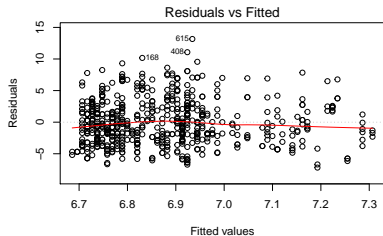
Linear model assumptions

- ▶ Linearity (transformations, GAM...)
- ▶ Residuals:
 - ▶ Independent
 - ▶ Equal variance
 - ▶ Normal

Linear model assumptions

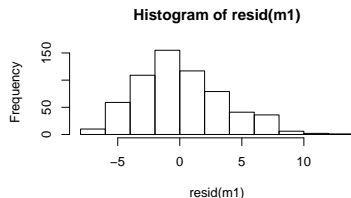
- ▶ Linearity (transformations, GAM...)
- ▶ Residuals:
 - ▶ Independent
 - ▶ Equal variance
 - ▶ Normal
- ▶ No measurement error in predictors

Model checking: residuals



Are residuals normal?

```
hist(resid(m1))
```



```
lm(formula = distance ~ age, data = paperplane)
      coef.est coef.se
(Intercept)  6.64    0.27
age           0.01    0.01
---
n = 615, k = 2
residual sd = 3.50, R-Squared = 0.00
```

SD of residuals = 3.5 coincides with estimate of σ .

How good is the model in predicting distance?

fitted gives predictions for each observation

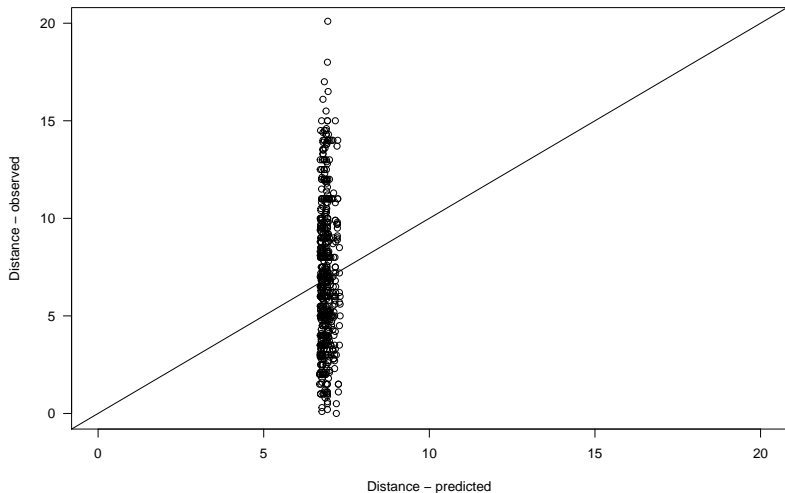
```
paperplanes$distance.pred <- fitted(m1)
head(paperplanes)
```

```
# A tibble: 6 x 9
```

	id	hour	person	gender	age	plane	paper	distance	distance.pred
	<int>	<fct>	<chr>	<fct>	<dbl>	<chr>	<int>	<dbl>	<dbl>
1	1	[17,18)	Roland	male	30.	Standa~	80	7.80	6.95
2	2	[17,18)	Astrid	female	30.	Concor~	120	2.70	6.95
3	3	[17,18)	Roland	male	30.	Standa~	120	9.20	6.95
4	4	[17,18)	Isabella	female	48.	Standa~	120	6.00	7.14
5	5	[17,18)	Fabienne	female	17.	Standa~	120	7.30	6.82
6	6	[17,18)	Fabienne	female	17.	Standa~	120	7.80	6.82

Calibration plot: Observed vs Predicted values

```
plot(paperplanes$distance.pred, paperplanes$distance, xlab = "Di
```



Using fitted model for prediction

Q: Expected distance if age = 30?

```
new.age <- data.frame(age = c(30))  
predict(m1, new.age, se.fit = TRUE)
```

```
$fit
```

```
1
```

```
6.954888
```

```
$se.fit
```

```
[1] 0.1633552
```

```
$df
```

```
[1] 613
```

```
$residual.scale
```

```
[1] 3.503736
```

Using fitted model for prediction

Q: Expected distance if age = 30?

```
new.age <- data.frame(age = c(30))  
predict(m1, new.age, se.fit = TRUE, interval = "confidence", lev
```

```
$fit  
      fit      lwr      upr  
1 6.954888 6.634085 7.275692
```

```
$se.fit  
[1] 0.1633552
```

```
$df  
[1] 613
```

```
$residual.scale  
[1] 3.503736
```


Using fitted model for prediction

Q: Expected distance if age = 30?

```
new.age <- data.frame(age = c(30))  
predict(m1, new.age, se.fit = TRUE, interval = "prediction", lev
```

```
$fit  
      fit      lwr      upr  
1 6.954888 0.06663211 13.84314
```

```
$se.fit  
[1] 0.1633552
```

```
$df  
[1] 613
```

```
$residual.scale  
[1] 3.503736
```

Important functions

- ▶ `plot`

Important functions

- ▶ `plot`
- ▶ `summary`

Important functions

- ▶ `plot`
- ▶ `summary`
- ▶ `coef`

Important functions

- ▶ `plot`
- ▶ `summary`
- ▶ `coef`
- ▶ `confint`

Important functions

- ▶ `plot`
- ▶ `summary`
- ▶ `coef`
- ▶ `confint`
- ▶ `fitted`

Important functions

- ▶ `plot`
- ▶ `summary`
- ▶ `coef`
- ▶ `confint`
- ▶ `fitted`
- ▶ `resid`

Important functions

- ▶ `plot`
- ▶ `summary`
- ▶ `coef`
- ▶ `confint`
- ▶ `fitted`
- ▶ `resid`
- ▶ `allEffects`

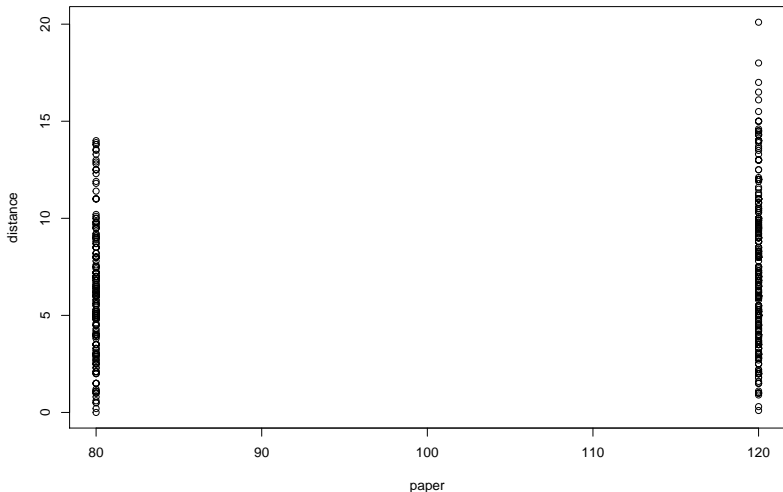
Important functions

- ▶ `plot`
- ▶ `summary`
- ▶ `coef`
- ▶ `confint`
- ▶ `fitted`
- ▶ `resid`
- ▶ `allEffects`
- ▶ `predict`

Categorical predictors (factors)

Q: Does distance vary with paper type?

```
plot(distance ~ paper, data = paperplanes)
```



Model distance ~ paper

All right here?

```
m2 <- lm(distance ~ paper, data = paperplanes)
```

Call:

```
lm(formula = distance ~ paper, data = paperplanes)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-7.2756	-2.3756	-0.3756	2.2244	12.7244

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	3.638290	0.750041	4.851	1.56e-06	***
paper	0.031144	0.007095	4.389	1.34e-05	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.453 on 613 degrees of freedom

Multiple R-squared: 0.03047, Adjusted R-squared: 0.02889

Model distance ~ paper

Paper is a factor!

```
paperplanes$paper <- as.factor(paperplanes$paper)
```

id	hour	person	gender
Min. : 1.0	[19,20) :139	Length:615	female:213
1st Qu.:154.5	[22,23) :108	Class :character	male :402
Median :308.0	[21,22) : 89	Mode :character	
Mean :308.0	[18,19) : 86		
3rd Qu.:461.5	[23,Inf): 78		
Max. :615.0	[17,18) : 75		
	(Other) : 40		

age	plane	paper	distance
Min. : 4.00	Length:615	80 :248	Min. : 0.000
1st Qu.:11.00	Class :character	120:367	1st Qu.: 4.350
Median :20.00	Mode :character		Median : 6.500
Mean :22.11			Mean : 6.873
3rd Qu.:28.00			3rd Qu.: 9.000
Max. :64.00			Max. :20.100

distance.pred

Model distance ~ paper

```
m2 <- lm(distance ~ paper, data = paperplanes)
```

Call:

```
lm(formula = distance ~ paper, data = paperplanes)
```

Residuals:

Min	1Q	Median	3Q	Max
-7.2756	-2.3756	-0.3756	2.2244	12.7244

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.1298	0.2192	27.958	< 2e-16 ***
paper120	1.2458	0.2838	4.389	1.34e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.453 on 613 degrees of freedom

Multiple R-squared: 0.03047, Adjusted R-squared: 0.02889

F-statistic: 19.27 on 1 and 613 DF, p-value: 1.339e-05

Linear model with categorical predictors

$$y_i = a + bx_i + \varepsilon_i$$

$$y_i = a + b_{paper120} + \varepsilon_i$$

Model distance ~ paper

```
m2 <- lm(distance ~ paper, data = paperplanes)
```

Call:

```
lm(formula = distance ~ paper, data = paperplanes)
```

Residuals:

Min	1Q	Median	3Q	Max
-7.2756	-2.3756	-0.3756	2.2244	12.7244

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.1298	0.2192	27.958	< 2e-16 ***
paper120	1.2458	0.2838	4.389	1.34e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.453 on 613 degrees of freedom

Multiple R-squared: 0.03047, Adjusted R-squared: 0.02889

F-statistic: 19.27 on 1 and 613 DF, p-value: 1.339e-05

Effects: Estimated Distance ~ paper

```
summary(allEffects(m2))
```

```
model: distance ~ paper
```

```
paper effect
```

```
paper
```

	80	120
--	----	-----

	6.129839	7.375613
--	----------	----------

```
Lower 95 Percent Confidence Limits
```

```
paper
```

	80	120
--	----	-----

	5.699269	7.021668
--	----------	----------

```
Upper 95 Percent Confidence Limits
```

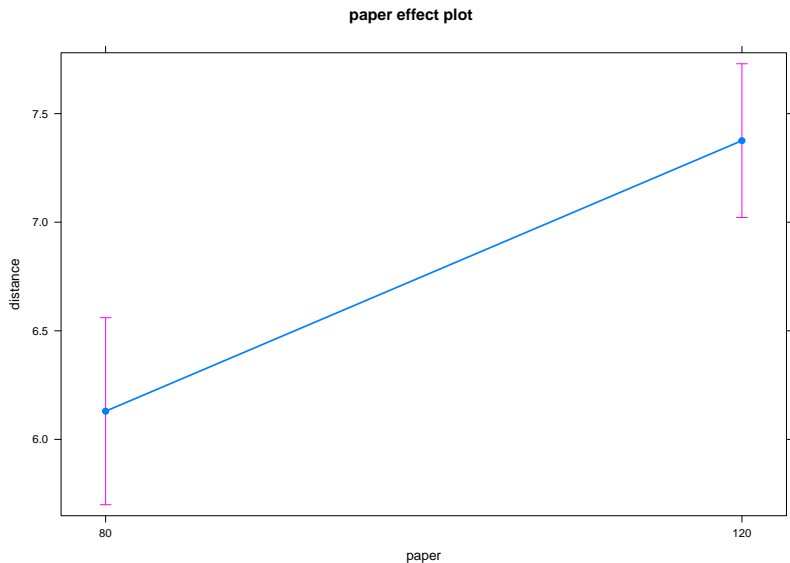
```
paper
```

	80	120
--	----	-----

	6.560408	7.729558
--	----------	----------

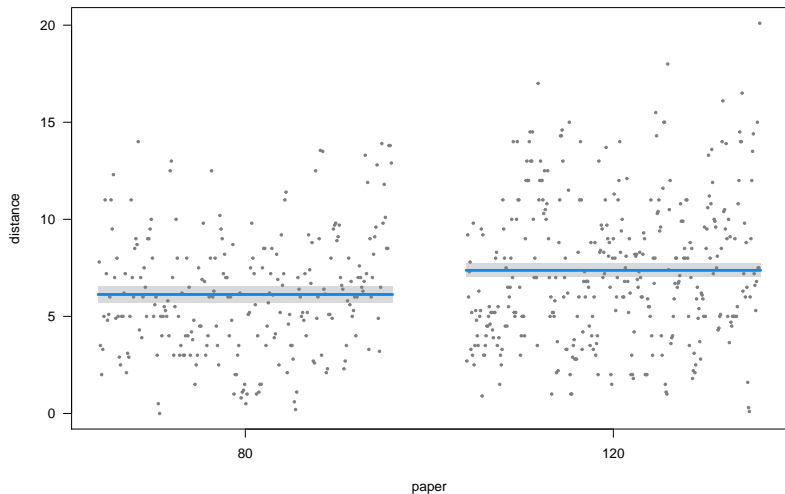
Plot

```
plot(allEffects(m2))
```

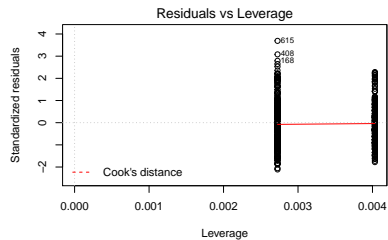
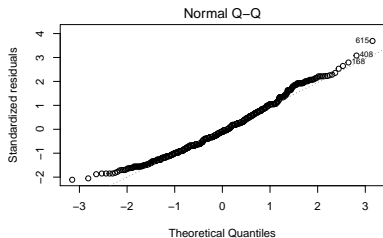
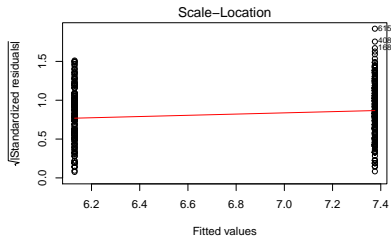
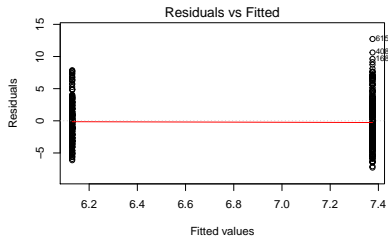


Plot (visreg)

```
visreg(m2)
```

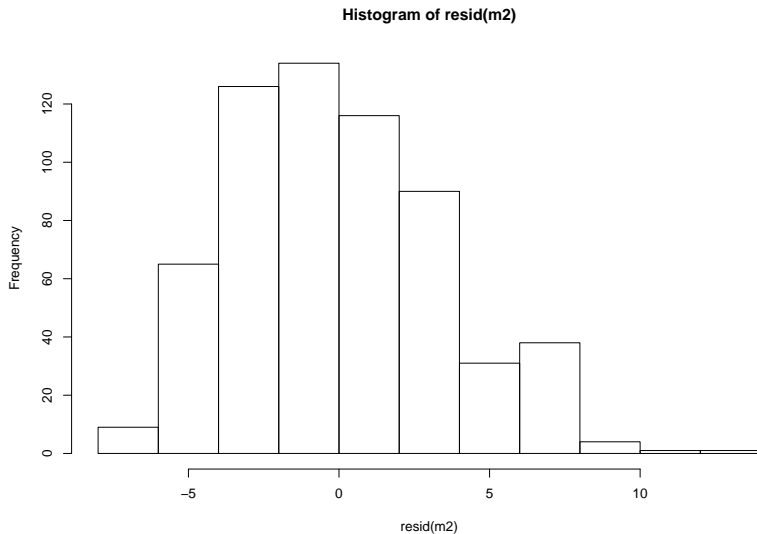


Model checking: residuals



Model checking: residuals

```
hist(resid(m2))
```



Exercise: Does distance vary with gender?

Combining continuous and categorical predictors

Predicting distance based on age and paper type

$$y_i = a + bx_i + \varepsilon_i$$

$$y_i = a + b_{paper120} + c \cdot age_i + \varepsilon_i$$

Predicting distance based on age and paper type

Call:

```
lm(formula = distance ~ age + paper, data = paperplanes)
```

Residuals:

Min	1Q	Median	3Q	Max
-7.1092	-2.4753	-0.3576	2.2523	12.5892

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.69210	0.33641	16.920	< 2e-16 ***
age	0.01774	0.01035	1.714	0.0871 .
paper120	1.32192	0.28683	4.609	4.93e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.447 on 612 degrees of freedom

Multiple R-squared: 0.0351, Adjusted R-squared: 0.03195

F-statistic: 11.13 on 2 and 612 DF, p-value: 1.784e-05

Estimated distance

```
summary(allEffects(multreg))
```

```
model: distance ~ age + paper
```

```
age effect
```

```
age
```

	4	20	30	50	60
	6.551921	6.835779	7.013191	7.368014	7.545425

```
Lower 95 Percent Confidence Limits
```

```
age
```

	4	20	30	50	60
	6.093516	6.559431	6.696578	6.738709	6.728156

```
Upper 95 Percent Confidence Limits
```

```
age
```

	4	20	30	50	60
	7.010326	7.112127	7.329803	7.997318	8.362694

```
paper effect
```

```
paper
```

	80	120
	6.084400	7.406318

```
Lower 95 Percent Confidence Limits
```

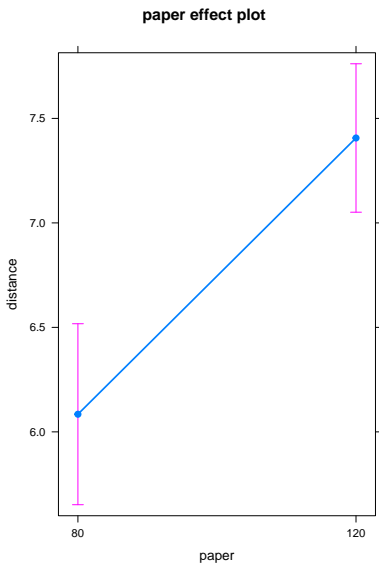
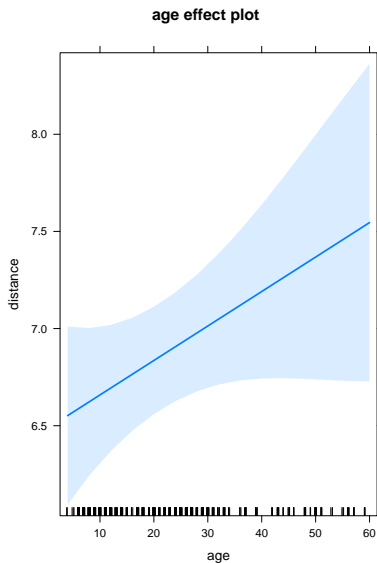
```
paper
```

	80	120
	5.651366	7.051182

```
Upper 95 Percent Confidence Limits
```

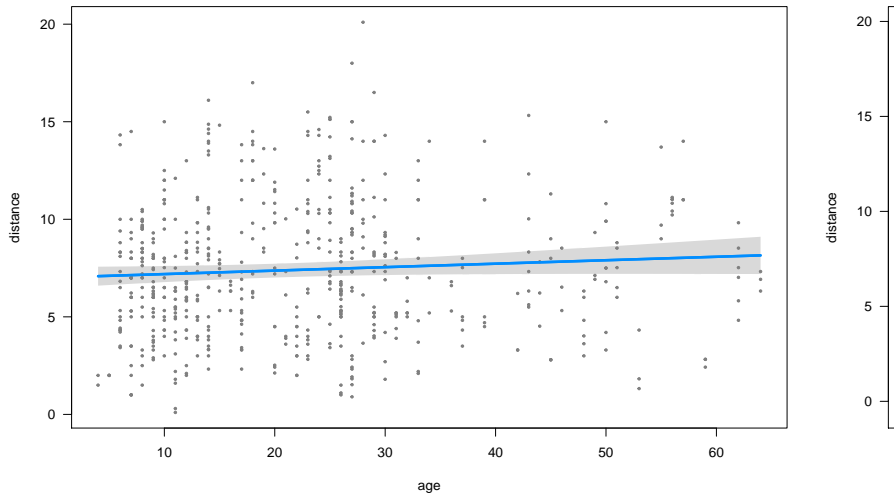
Plot

```
plot(allEffects(multreg))
```

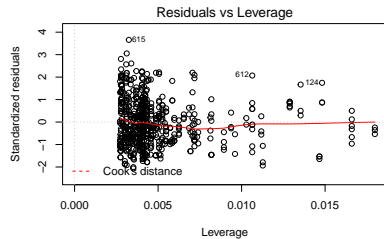
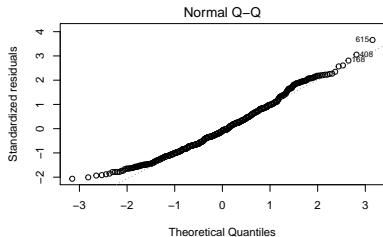
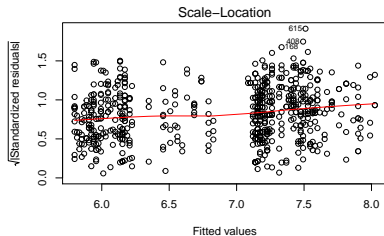
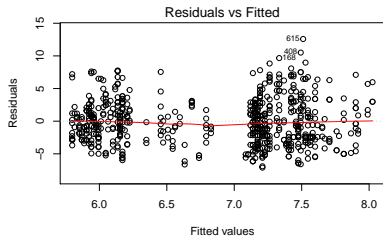


Plot (visreg)

```
visreg(multreg)
```

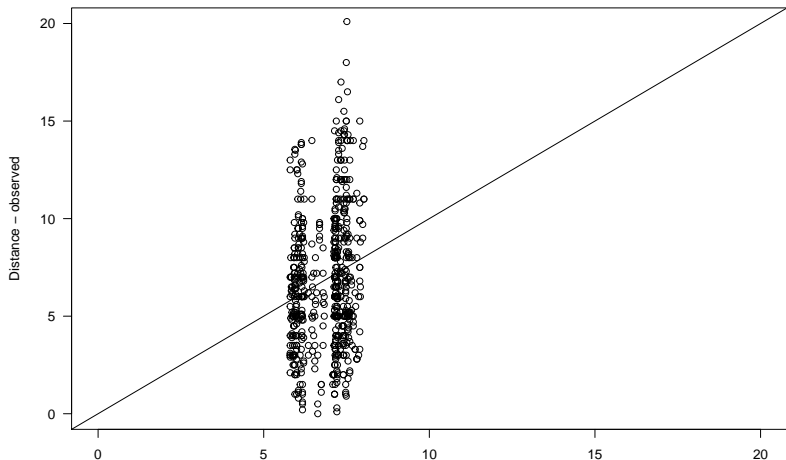


Model checking: residuals



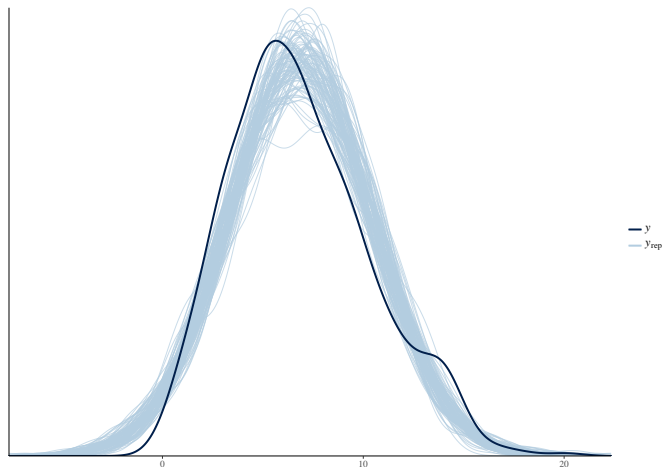
How good is this model? Calibration plot

```
paperplanes$distance.pred <- fitted(multreg)
plot(paperplanes$distance.pred, paperplanes$distance, xlab = "Di
abline(a = 0, b = 1)
```



Model checking with simulated data

```
library(bayesplot)
sims <- simulate(multreg, nsim = 100)
ppc_dens_overlay(paperplanes$distance, yrep = t(as.matrix(sims)))
```



Extra exercises

- ▶ mammal sleep: Are sleep patterns related to diet?

Extra exercises

- ▶ mammal sleep: Are sleep patterns related to diet?
- ▶ iris: Predict petal length \sim petal width and species