

Linear models

Example dataset: forest trees

- Go to <https://tinyurl.com/treesdata>

```
trees <- read.csv("data-raw/trees.csv")  
head(trees)
```

	plot	dbh	height	sex	dead
1	4	29.68	36.1	male	0
2	5	33.29	42.3	male	0
3	2	28.03	41.9	female	0
4	5	39.86	46.5	female	0
5	1	47.94	43.9	female	0
6	1	10.82	26.2	male	0

Example dataset: forest trees

- ▶ Go to <https://tinyurl.com/treesdata>
- ▶ Download zip file and uncompress (within your project folder!)

```
trees <- read.csv("data-raw/trees.csv")  
head(trees)
```

	plot	dbh	height	sex	dead
1	4	29.68	36.1	male	0
2	5	33.29	42.3	male	0
3	2	28.03	41.9	female	0
4	5	39.86	46.5	female	0
5	1	47.94	43.9	female	0
6	1	10.82	26.2	male	0

Questions

- ▶ What is the relationship between DBH and height?

Questions

- ▶ What is the relationship between DBH and height?
- ▶ Do taller trees have bigger trunks?

Questions

- ▶ What is the relationship between DBH and height?
- ▶ Do taller trees have bigger trunks?
- ▶ Can we predict height from DBH? How well?

Always plot your data first!

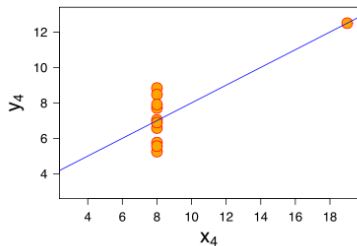
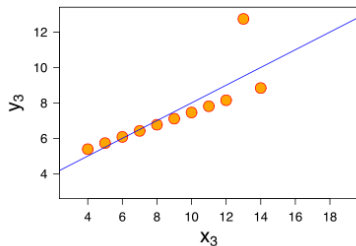
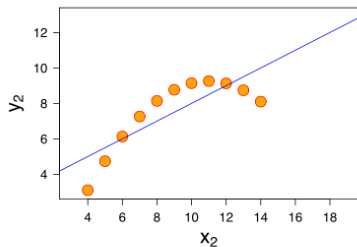
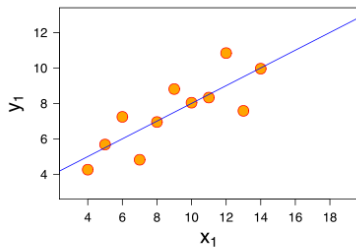
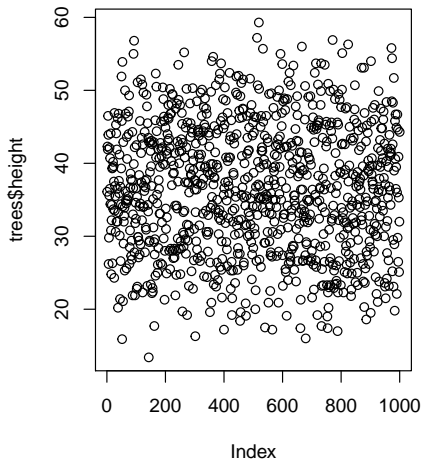


Figure 1:

Exploratory Data Analysis (EDA)

Outliers

```
plot(trees$height)
```



Outliers impact on regression

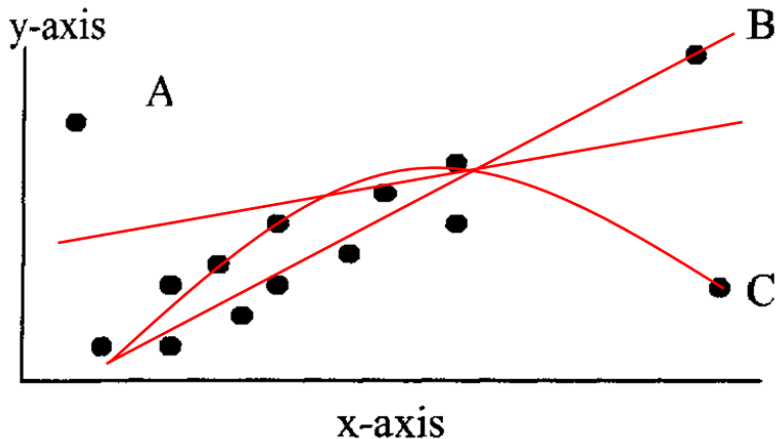
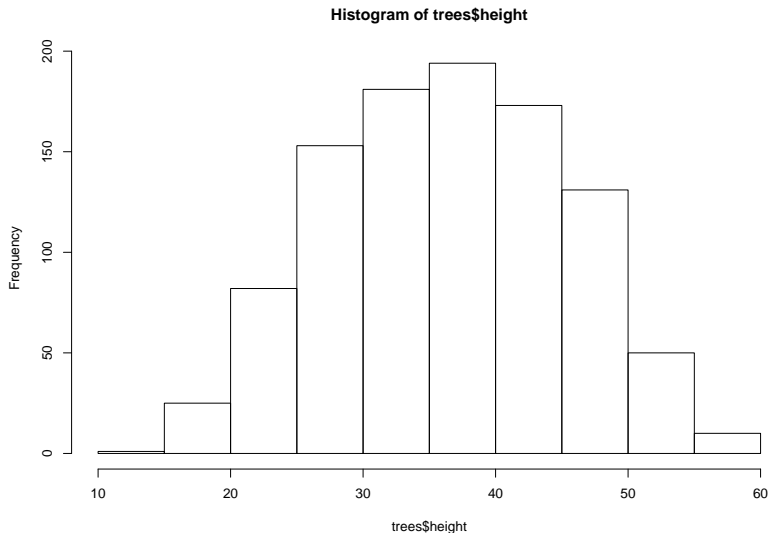


Figure 2:

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

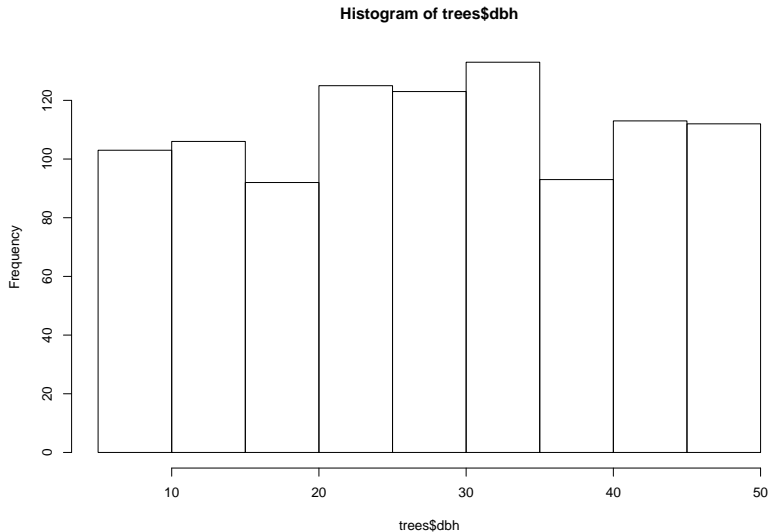
Histogram of response variable

```
hist(trees$height)
```



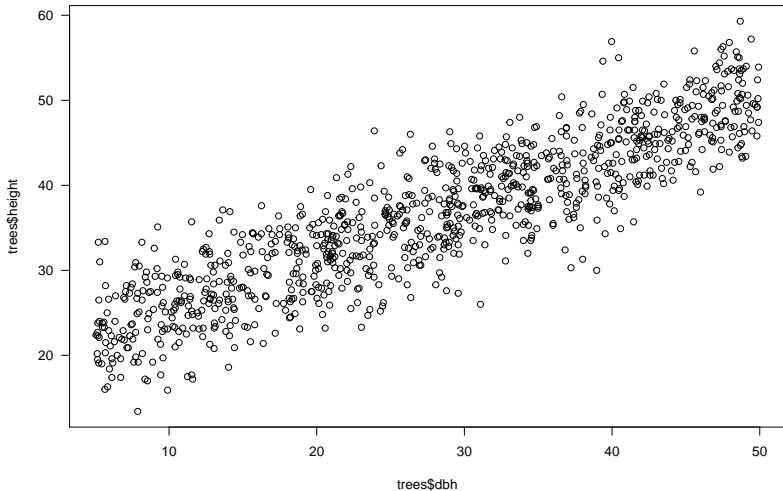
Histogram of predictor variable

```
hist(trees$dbh)
```



Scatterplot

```
plot(trees$dbh, trees$height, las = 1)
```



Now fit model

Hint: `lm`

Now fit model

Hint: `lm`

```
m1 <- lm(height ~ dbh, data = trees)
```

What does this mean?

Call:

```
lm(formula = height ~ dbh, data = trees)
```

Residuals:

Min	1Q	Median	3Q	Max
-13.3270	-2.8978	0.1057	2.7924	12.9511

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	19.33920	0.31064	62.26	<2e-16 ***
dbh	0.61570	0.01013	60.79	<2e-16 ***

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

Residual standard error: 4.093 on 998 degrees of freedom

Multiple R-squared: 0.7874, Adjusted R-squared: 0.7871

F-statistic: 3695 on 1 and 998 DF, p-value: < 2.2e-16

Retrieving model coefficients

```
coef(m1)
```

(Intercept)	dbh
19.3391968	0.6157036

Tidy up model coefficients with broom

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

	term	estimate	std.error	statistic	p.value
1	(Intercept)	19.3391968	0.31064458	62.25506	0
2	dbh	0.6157036	0.01012841	60.78976	0

```
glance(m1)
```

	r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik
1	0.7873608	0.7871477	4.092629	3695.395	0	2	-2827.12

	BIC	deviance	df.residual
1	5674.973	16716.11	998

Confidence intervals

```
confint(m1)
```

	2.5 %	97.5 %
(Intercept)	18.7296053	19.948788
dbh	0.5958282	0.635579

Using effects package

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

model: height ~ dbh

dbh effect

dbh

	5	20	30	40	50
	22.41771	31.65327	37.81030	43.96734	50.12438

Lower 95 Percent Confidence Limits

dbh

	5	20	30	40	50
	21.89682	31.35487	37.55287	43.61733	49.61669

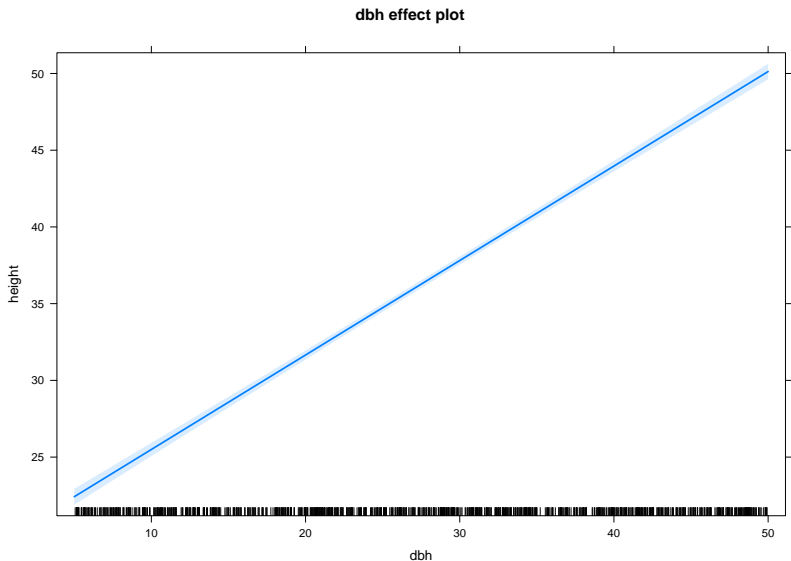
Upper 95 Percent Confidence Limits

dbh

	5	20	30	40	50
	22.93861	31.95167	38.06774	44.31735	50.63207

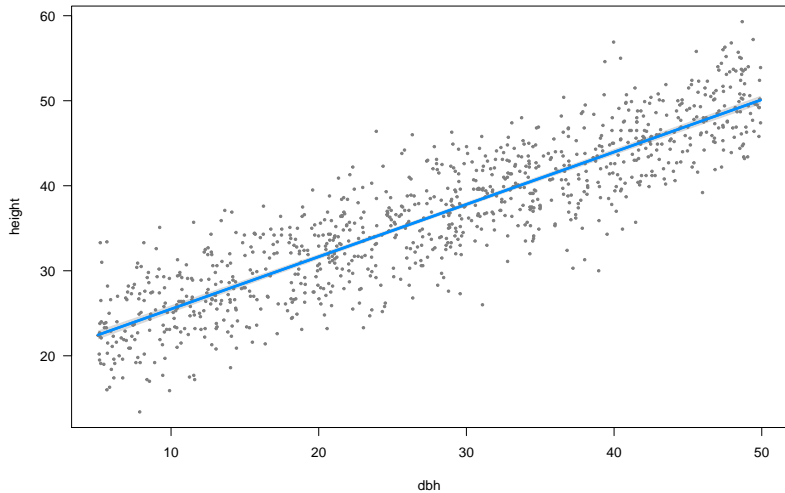
Plot effects

```
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

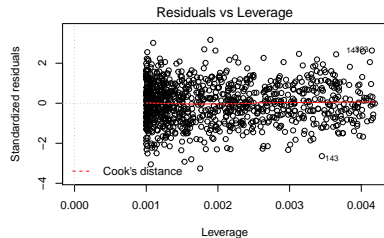
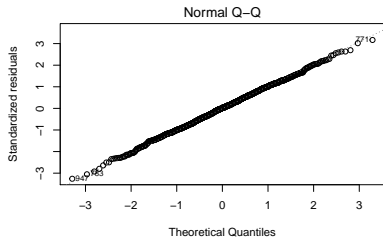
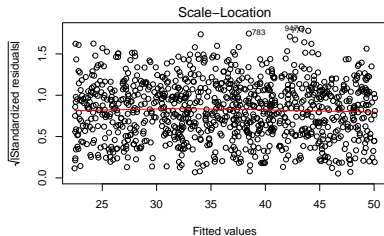
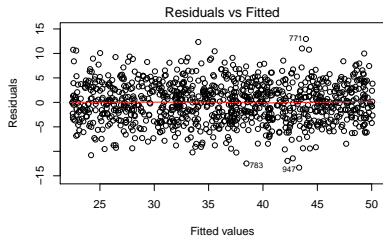
Linear model assumptions

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

Linear model assumptions

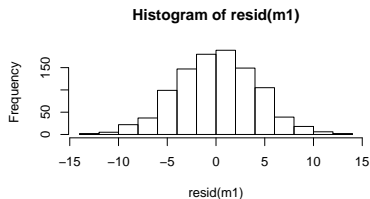
- ▶ Linearity (transformations, GAM...)
- ▶ Residuals:
 - ▶ Independent
 - ▶ Equal variance
 - ▶ Normal
- ▶ No measurement error in predictors

Model checking: residuals



Are residuals normal?

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



```
lm(formula = height ~ dbh, data = trees)
```

	coef.est	coef.se
(Intercept)	19.34	0.31
dbh	0.62	0.01

n = 1000, k = 2

residual sd = 4.09, R-Squared = 0.79

SD of residuals = 4.09 coincides with estimate of σ .

How good is the model in predicting tree height?

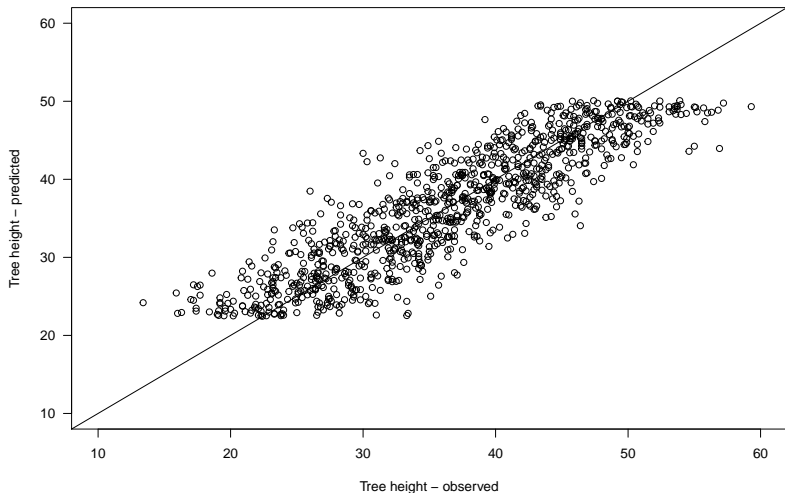
fitted gives predictions for each observation

```
trees$height.pred <- fitted(m1)  
head(trees)
```

	plot	dbh	height	sex	dead	height.pred
1	4	29.68	36.1	male	0	37.61328
2	5	33.29	42.3	male	0	39.83597
3	2	28.03	41.9	female	0	36.59737
4	5	39.86	46.5	female	0	43.88114
5	1	47.94	43.9	female	0	48.85603
6	1	10.82	26.2	male	0	26.00111

Calibration plot: Observed vs Predicted values

```
plot(trees$height, trees$height.pred, xlab = "Tree height - obse
```



Using fitted model for prediction

Q: Expected tree height if DBH = 39 cm?

```
new.dbh <- data.frame(dbh = c(39))  
predict(m1, new.dbh, se.fit = TRUE)
```

```
$fit
```

```
1
```

```
43.35164
```

```
$se.fit
```

```
[1] 0.1715514
```

```
$df
```

```
[1] 998
```

```
$residual.scale
```

```
[1] 4.092629
```


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`

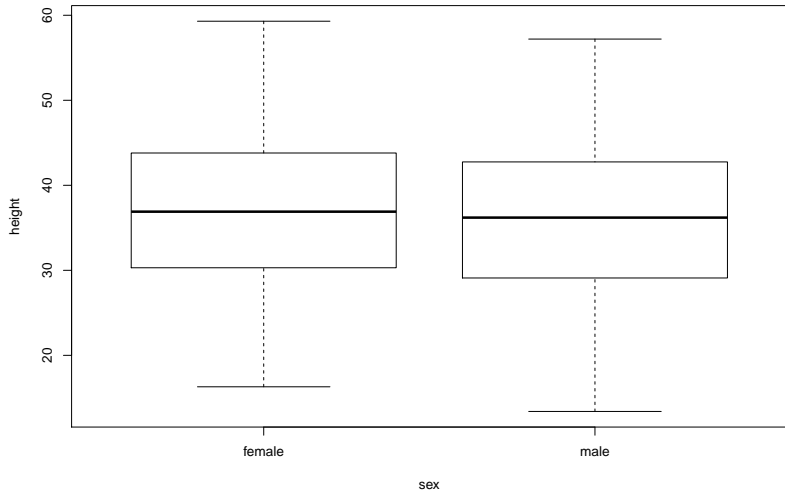
Important functions

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

Categorical predictors (factors)

Q: Does tree height vary with sex?

```
plot(height ~ sex, data = trees)
```



Model height ~ sex

```
m2 <- lm(height ~ sex, data = trees)
```

Call:

```
lm(formula = height ~ sex, data = trees)
```

Residuals:

Min	1Q	Median	3Q	Max
-22.6881	-6.7881	-0.0097	6.7261	22.3687

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	36.9312	0.3981	92.778	<2e-16 ***
sexmale	-0.8432	0.5607	-1.504	0.133

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

Residual standard error: 8.865 on 998 degrees of freedom

Multiple R-squared: 0.002261, Adjusted R-squared: 0.001261

F-statistic: 2.261 on 1 and 998 DF, p-value: 0.133

Linear model with categorical predictors

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

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

Model height ~ sex

```
m2 <- lm(height ~ sex, data = trees)
```

Call:

```
lm(formula = height ~ sex, data = trees)
```

Residuals:

Min	1Q	Median	3Q	Max
-22.6881	-6.7881	-0.0097	6.7261	22.3687

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	36.9312	0.3981	92.778	<2e-16 ***
sexmale	-0.8432	0.5607	-1.504	0.133

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

Residual standard error: 8.865 on 998 degrees of freedom

Multiple R-squared: 0.002261, Adjusted R-squared: 0.001261

F-statistic: 2.261 on 1 and 998 DF, p-value: 0.133

Effects: Height ~ sex

Compare CIs

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

```
model: height ~ sex
```

```
sex effect
```

```
sex
```

```
female      male
```

```
36.93125 36.08810
```

```
Lower 95 Percent Confidence Limits
```

```
sex
```

```
female      male
```

```
36.15012 35.31319
```

```
Upper 95 Percent Confidence Limits
```

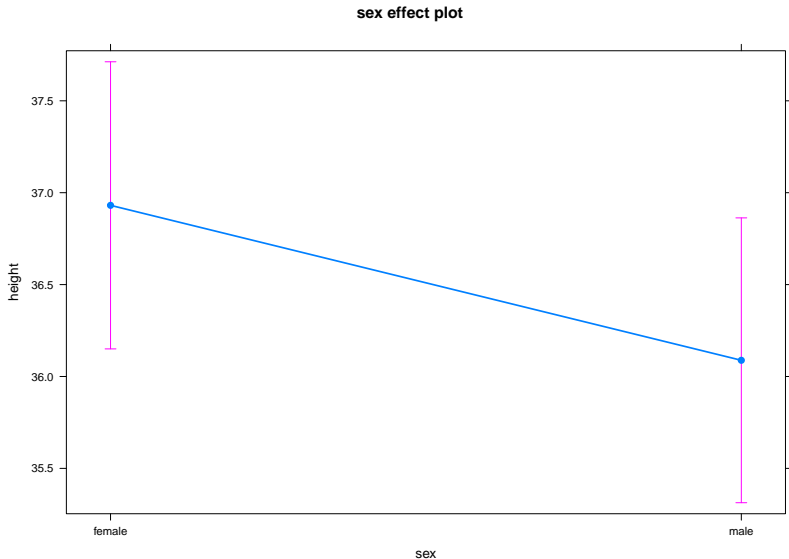
```
sex
```

```
female      male
```

```
37.71238 36.86300
```

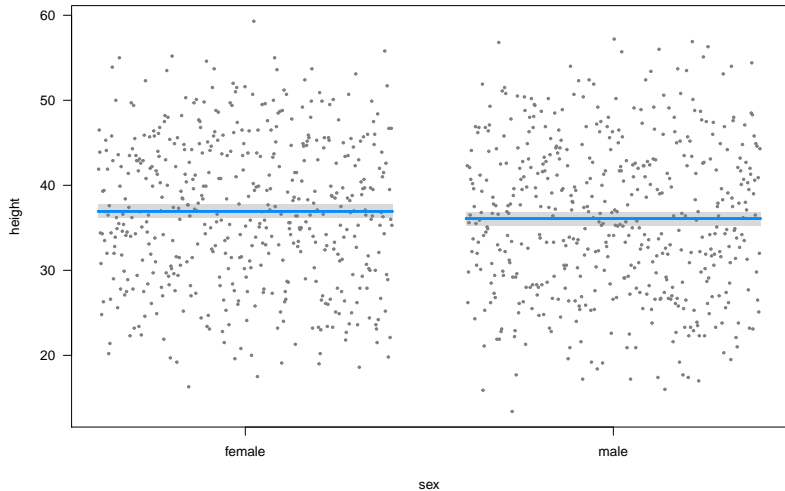
Plot

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

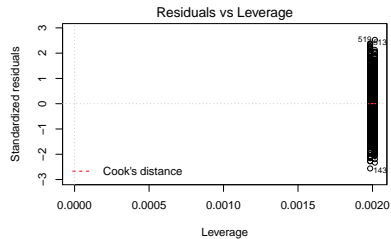
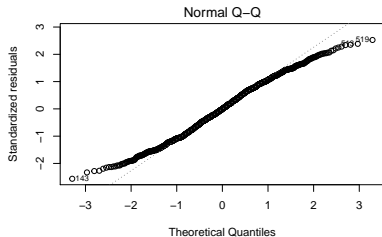
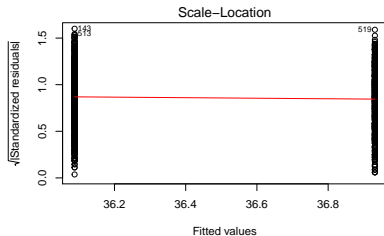
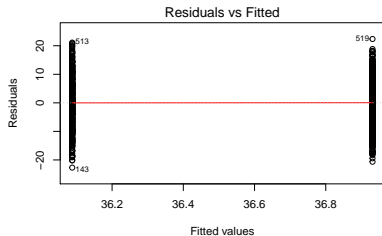


Plot (visreg)

```
visreg(m2)
```

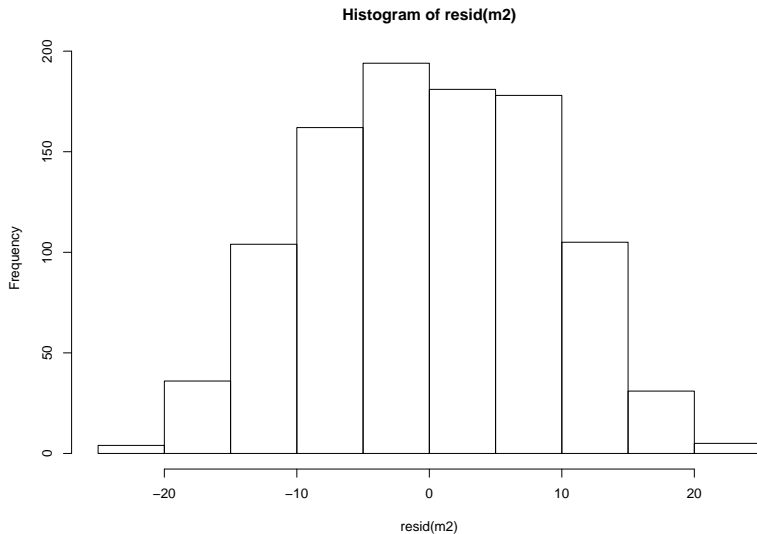


Model checking: residuals



Model checking: residuals

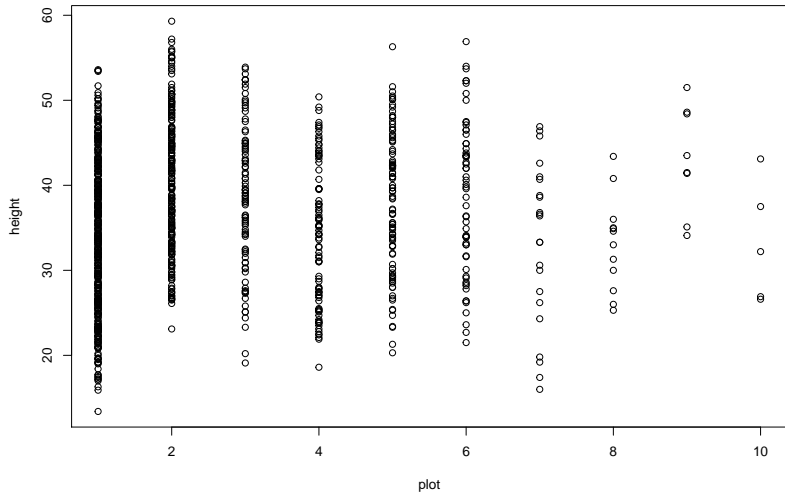
```
hist(resid(m2))
```



Q: Does height differ among field plots?

Plot data first

```
plot(height ~ plot, data = trees)
```



Linear model with categorical predictors

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

$$y_i = a + b_{plot2} + c_{plot3} + d_{plot4} + e_{plot5} + \dots + \varepsilon_i$$

Model Height ~ Plot

All right here?

```
m3 <- lm(height ~ plot, data = trees)
```

Call:

```
lm(formula = height ~ plot, data = trees)
```

Residuals:

Min	1Q	Median	3Q	Max
-22.4498	-6.7049	0.0709	6.7537	23.0640

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	35.4636	0.4730	74.975	< 2e-16 ***
plot	0.3862	0.1413	2.733	0.00639 **

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

Residual standard error: 8.842 on 998 degrees of freedom

Multiple R-squared: 0.007429, Adjusted R-squared: 0.006435

Plot is a factor!

```
trees$plot <- as.factor(trees$plot)
```

Model Height ~ Plot

Call:

```
lm(formula = height ~ plot, data = trees)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-20.4416	-6.9004	0.0379	6.3051	19.7584

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	33.8416	0.4266	79.329	< 2e-16	***
plot2	6.3411	0.7126	8.899	< 2e-16	***
plot3	4.9991	0.9828	5.086	4.36e-07	***
plot4	0.5329	0.9872	0.540	0.58949	
plot5	4.3723	0.9425	4.639	3.97e-06	***
plot6	4.7601	1.1709	4.065	5.18e-05	***
plot7	-0.7416	1.8506	-0.401	0.68871	
plot8	-0.6832	2.4753	-0.276	0.78258	
plot9	9.1709	3.0165	3.040	0.00243	**
plot10	-0.5816	3.8013	-0.153	0.87843	

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

Residual standard error: 8.446 on 990 degrees of freedom

Multiple R-squared: 0.1016, Adjusted R-squared: 0.09344

F-statistic: 12.44 on 9 and 990 DF, p-value: < 2.2e-16

Estimated tree heights for each site

```
summary(allEffects(m3))
```

```
model: height ~ plot
```

```
plot effect
```

```
plot
```

	1	2	3	4	5	6	7	8
	33.84158	40.18265	38.84066	34.37444	38.21386	38.60167	33.10000	33.15833
	9	10						
	43.01250	33.26000						

```
Lower 95 Percent Confidence Limits
```

```
plot
```

	1	2	3	4	5	6	7	8
	33.00444	39.06264	37.10317	32.62733	36.56463	36.46190	29.56629	28.37367
	9	10						
	37.15251	25.84764						

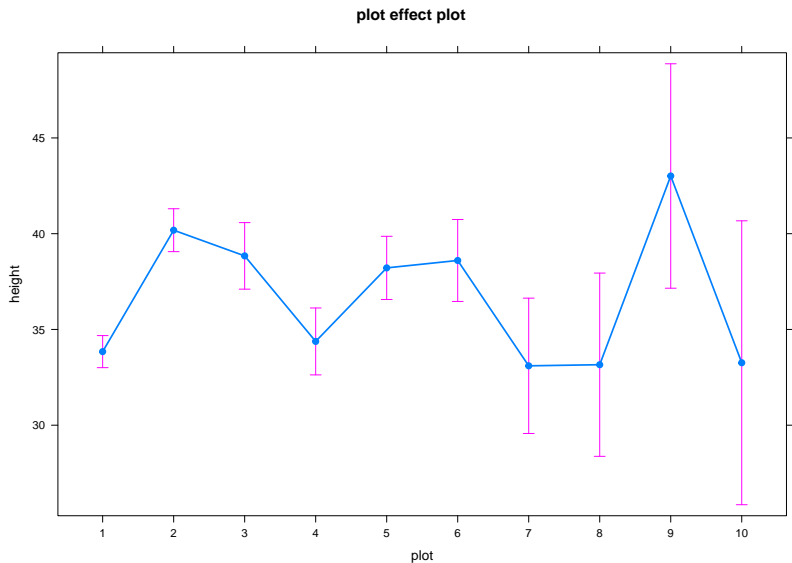
```
Upper 95 Percent Confidence Limits
```

```
plot
```

	1	2	3	4	5	6	7	8
	34.67872	41.30265	40.57814	36.12156	39.86309	40.74143	36.63371	37.94299
	9	10						
	48.87249	40.67236						

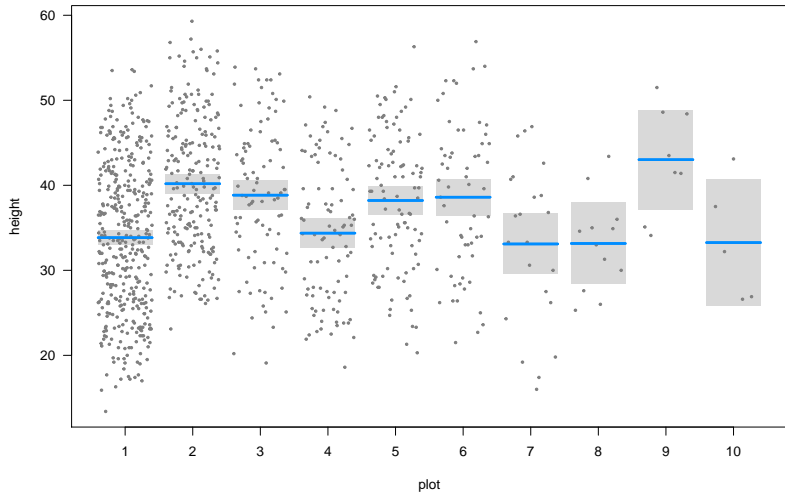
Plot

```
plot(allEffects(m3))
```

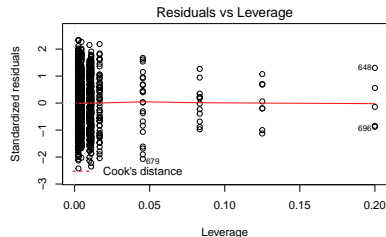
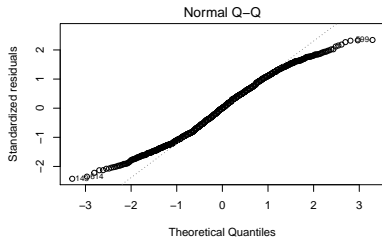
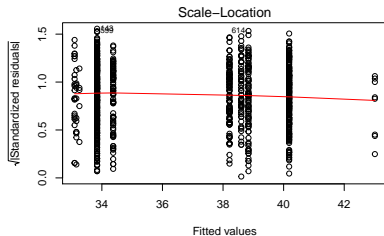
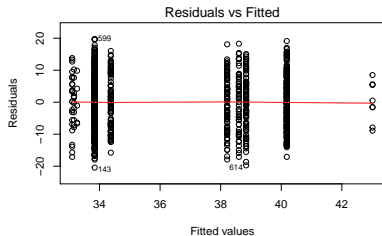


Plot (visreg)

```
visreg(m3)
```



Model checking: residuals



Combining continuous and categorical predictors

Predicting tree height based on dbh and site

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

$$y_i = a + b_{plot2} + c_{plot3} + d_{plot4} + e_{plot5} + \dots + k \cdot DBH_i + \varepsilon_i$$

Predicting tree height based on dbh and site

Call:

```
lm(formula = height ~ plot + dbh, data = trees)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-10.1130	-1.9885	0.0582	2.0314	11.3320

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	16.699037	0.260565	64.088	< 2e-16 ***
plot2	6.504303	0.256730	25.335	< 2e-16 ***
plot3	4.357457	0.354181	12.303	< 2e-16 ***
plot4	1.934650	0.356102	5.433	6.98e-08 ***
plot5	3.637432	0.339688	10.708	< 2e-16 ***
plot6	4.204511	0.421906	9.966	< 2e-16 ***
plot7	-0.176193	0.666772	-0.264	0.7916
plot8	-5.312648	0.893603	-5.945	3.82e-09 ***
plot9	5.437049	1.087766	4.998	6.84e-07 ***
plot10	2.263338	1.369986	1.652	0.0988 .
dbh	0.617075	0.007574	81.473	< 2e-16 ***

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

Residual standard error: 3.043 on 989 degrees of freedom

Multiple R-squared: 0.8835, Adjusted R-squared: 0.8823

F-statistic: 750 on 10 and 989 DF, p-value: < 2.2e-16

Estimated tree heights for each site

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

```
model: height ~ plot + dbh
```

```
plot effect
```

```
plot
```

	1	2	3	4	5	6	7	8
	33.90437	40.40868	38.26183	35.83902	37.54181	38.10889	33.72818	28.59173
	9	10						
	39.34142	36.16771						

```
Lower 95 Percent Confidence Limits
```

```
plot
```

	1	2	3	4	5	6	7	8
	33.60276	40.00512	37.63569	35.20858	36.94739	37.33787	32.45495	26.86438
	9	10						
	37.22831	33.49623						

```
Upper 95 Percent Confidence Limits
```

```
plot
```

	1	2	3	4	5	6	7	8
	34.20599	40.81223	38.88798	36.46947	38.13622	38.87990	35.00141	30.31907
	9	10						
	41.45454	38.83919						

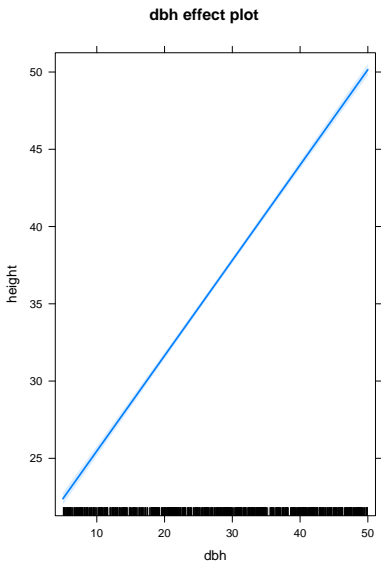
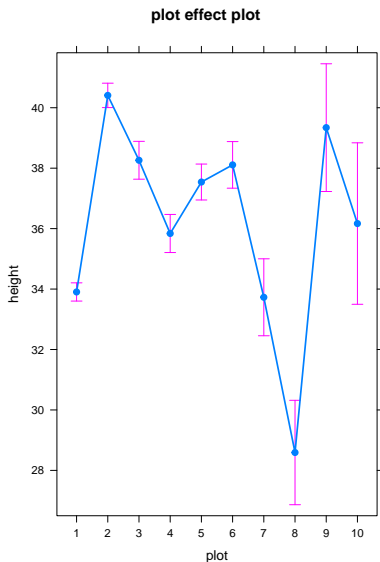
```
dbh effect
```

```
dbh
```

	5	20	30	40	50
	22.38634	31.64246	37.81321	43.98396	50.15471

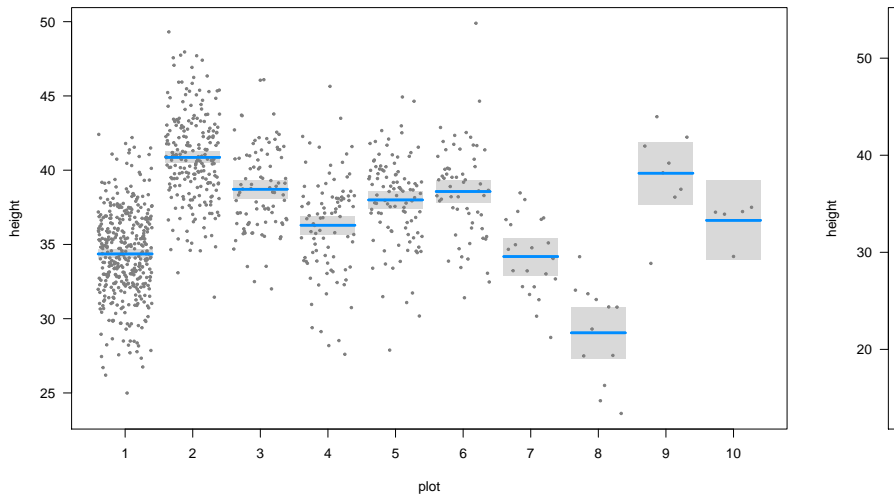
Plot

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

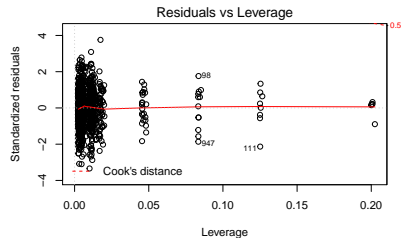
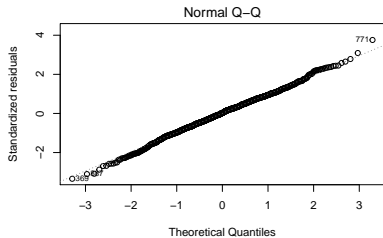
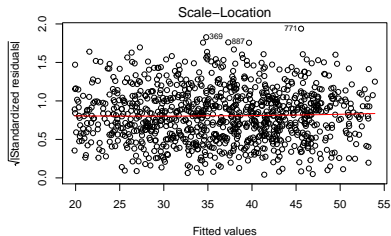
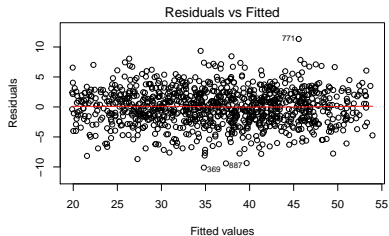


Plot (visreg)

```
visreg(multreg)
```

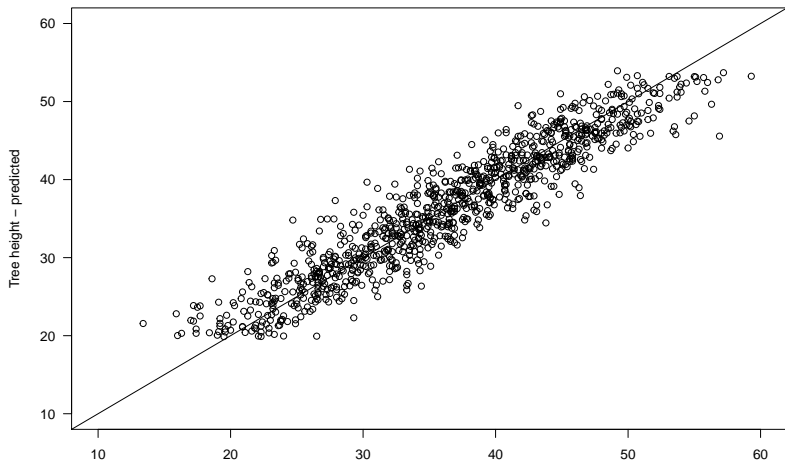


Model checking: residuals



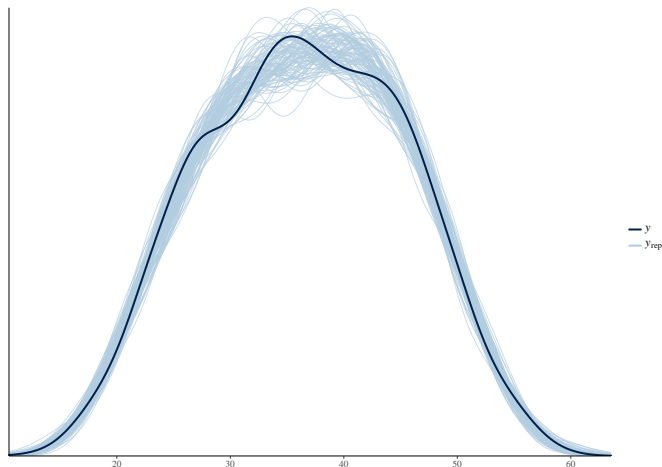
How good is this model? Calibration plot

```
trees$height.pred <- fitted(multreg)
plot(trees$height, trees$height.pred, xlab = "Tree height - obse
abline(a = 0, b = 1)
```



Model checking with simulated data

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



Extra exercises

- ▶ paperplanes: How does flight distance differ with age, gender or paper type?

Extra exercises

- ▶ paperplanes: How does flight distance differ with age, gender or paper type?
- ▶ mammal sleep: Are sleep patterns related to diet?

Extra exercises

- ▶ paperplanes: How does flight distance differ with age, gender or paper type?
- ▶ mammal sleep: Are sleep patterns related to diet?
- ▶ iris: Predict petal length \sim petal width and species