Regression (Chapter 11)

Setup: Have data on two *quantitative* variables — X and Y — on a sample of n subjects.

Q: Is there any association between X and Y? What kind?

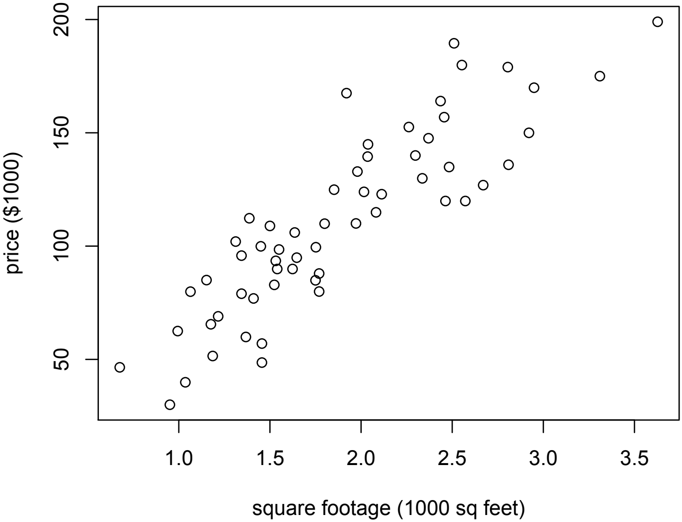
Scatterplot:

- Plot y against x
- Look for the **trend** in the plot — a smooth curve that shows how the average value of Y changes with x
- Trend may be linear or non-linear
- If there is a trend, then the two variables are associated. In this case, x may be used to predict y
- Trend may be strong or weak. It is strong if the points are tightly clustered around the trend (small scatter)
- No trend: No association — i.e., the variables are independent, and x is not helpful for predicting y.

Example: House price data

```
house <- read.table(file="house_price.txt", sep=",",
header=T)

> head(house)
   size price
1 0.951 30.00
2 1.036 39.90
3 0.676 46.50
4 1.456 48.60
5 1.186 51.50
6 1.456 56.99
>

> str(house)
'data.frame': 58 obs. of  2 variables:
 $ size : num  0.951 1.036 0.676 1.456 1.186 ...
```

```
 $ price: num  30 39.9 46.5 48.6 51.5 ...
>

# Make a scatterplot

plot(house$size, house$price,
xlab="square footage (1000 sq feet)",
ylab="price ($1000)")
```


price ($1000)
200
150
100
50
1.0
1.5
2.0
2.5
3.0
3.5
square footage (1000 sq feet)

Overall pattern in a scatterplot

Form:

Direction:

Strength - assess the scatter of the points:

Outliers - observations that don't follow overall pattern:

Correlation Coefficient

Population correlation: A measure of **linear** relationship between X and Y. It is defined as,

$$\rho = \frac{\text{cov}(X,Y)}{\text{sd}(X)\text{sd}(Y)},$$

where $\text{cov}(X,Y) = E\{(X - E(X))(Y - E(Y))\}$ is covariance between X and Y.

Sample correlation: Estimator of ρ. It is defined as

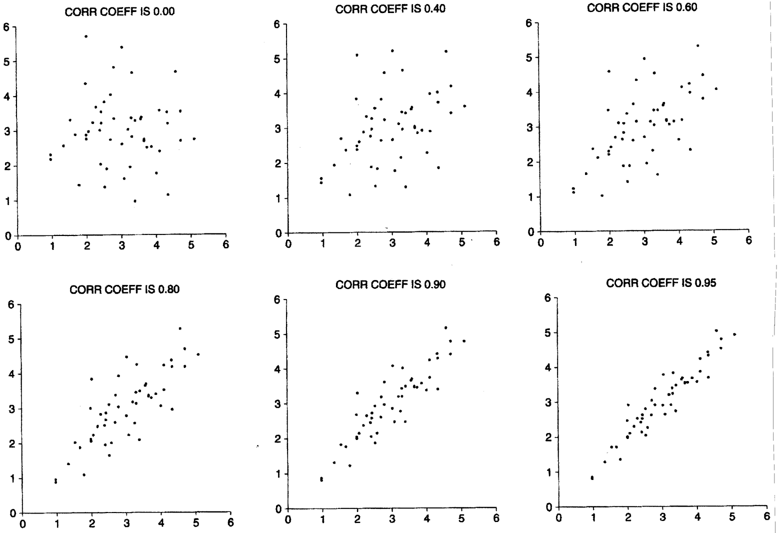
$$r = \frac{S_{xy}}{S_x S_y},$$

where

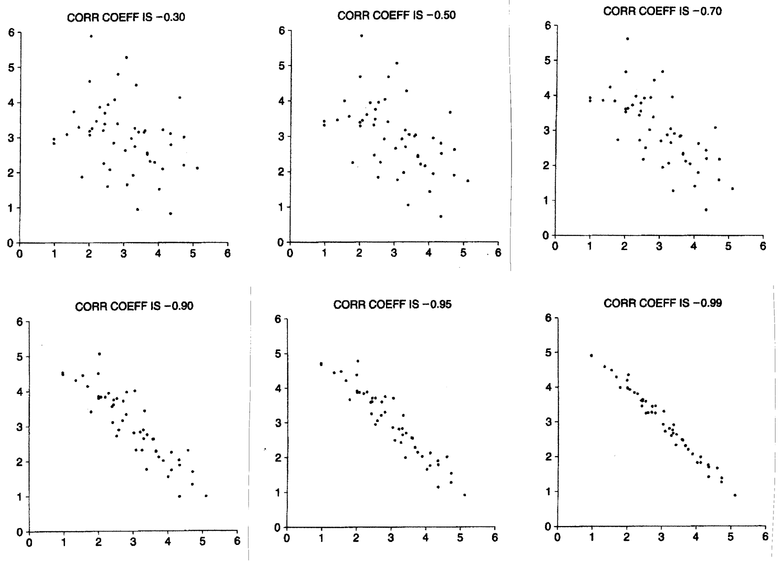
Properties of ρ and r:

- Range between -1 to 1
- Sign tells us
- Absolute value tells us
- Perfect correlation:
- Unit free
- **No change** if X and Y are interchanged or if X is replaced by $aX + b$ and/or Y is replaced by $cY + d$, where a and c have the same sign. The sign will reverse if a and c have different signs.
- Zero correlation: **No linear** relationship. But there may be non-linear relationship.
- Independent X and Y: zero correlation, but the converse may not be true

Examples of Positive Correlation



Examples of Negative Correlation



Caution: Non-linear relationships

Ex: Scatter plot of speed and mileage (miles per gallon) of an automobile.

More non-linear patterns:

You can always compute r — but it doesn't makes sense for curve patters.
Lesson: Correlation only measures the strength of **linear** association — i.e., how close the points are to a straight line.

Caution: Outliers

The position of an outlier relative to the rest of the ("cloud") of points determines how it affects r. Outliers may decrease or increase the value of r.

Note: Just knowing the value of r will give no information about whether outliers are present. That's why it is important to look at scatterplots.

Simple Linear Regression

Setup: Have data $(X_i, Y_i), i = 1, \ldots, n$, on two quantitative variables X & Y. Their scatterplot shows a linear relationship. Need an equation that would allow us to predict Y from X.

Response variable (Y): variable to be predicted (or modeled), aka, *dependent variable*.

Predictor (X): variable used to predict Y, aka, *independent or explanatory variable or covariate*.

Regression model: A function that models mean response — $E(Y|X = x)$ — as a function of x

Simple linear regression: $E(Y|X = x) = \beta_0 + \beta_1 x$

- Assumes mean response changes *linearly* with x
- β_0 : intercept — $E(Y|X = 0)$
- β_1 : slope — rate of change of mean response. It represents the change in mean when x increases by 1 unit.

- The regression coefficients are estimated from data.
- Let $(\hat{\beta}_0, \hat{\beta}_1)$ = estimator of (β_0, β_1).

Observed response: Y_i when $X = x_i$, $i = 1, \ldots, n$.

Fitted response: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$

- *Estimated mean response* when $X = x$
- *Response predicted* by the regression line
- $(\hat{Y}, x)$ falls on the regression line
- Fitted values: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $i = 1, \ldots, n$.

Residuals: $e_i = Y_i - \hat{Y}_i$, $i = 1, \ldots, n$.

- Vertical distance between observed and predicted Y's
- Error in prediction
- Large residuals: observed and fitted Ys are too far

Least squares method for estimating coefficients: Find $(\hat{\beta}_0, \hat{\beta}_1)$ that minimize the *sum of squares of residuals*

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

- Results in the **line of best fit** — the line is such that the fitted Ys are "closest" to the observed Ys
- Fitted regression line: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- Other criteria possible, e.g., minimizing $\sum_{i=1}^{n} |e_i|$, but the resulting estimates don't have simple expressions

To minimize $\sum_{i=1}^{n} e_i^2$ wrt (β_0, β_1), solve the **normal equations**

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \beta_0} = 0, \quad \frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \beta_1} = 0,$$

resulting in the **least squares estimates**

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}, \quad \hat{\beta}_1 = r S_y / S_x,$$

where r is **sample correlation**, and S_x and S_y are **standard deviations** of x and y samples, respectively.

Recall that:

$$S_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2, \quad S_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2,$$

$$r = \frac{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{S_x S_y}$$

The fitted regression line

Fitted regression line: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Plugging-in $\hat{\beta}_0$ and $\hat{\beta}_1$,

$$\hat{Y} =$$

$$=$$

implying that

$$\frac{\hat{Y} - \overline{Y}}{S_y} = r\frac{(x - \overline{x})}{S_x}.$$

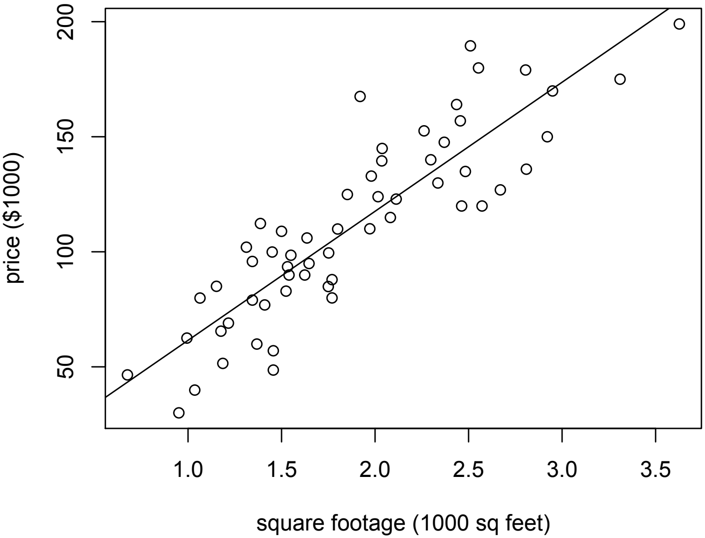
- If x is 1 SD away from its mean $\overline{x}$, $\hat{Y}$ is r SD away from its mean $\overline{Y}$. Since $|r| \leq 1$, this means $\hat{Y}$ is **closer** to $\overline{Y}$ (in units of SD) than x is to $\overline{x}$ — **regression toward mean**.
- The fitted line passes through the points $(\overline{x}, \overline{y})$.
- The sign of slope $\hat{\beta}_1$ is same as the sign of r.
- The sum of residuals, $\sum_{i=1}^{n} e_i =$
- The average of fitted values, $(1/n)\sum_{i=1}^{n} \hat{Y}_i =$

Ex: Let's get the fitted line for the house price data and add it to the scatterplot.

```
x <- house$size
y <- house$price

# Get the fitted regression line
> (house.reg <- lm (y ~ x))
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
      5.432       56.083
>

# Does R do what we expect it to do?
> c(mean(x), sd(x), mean(y), sd(y), cor(x,y))
```

```
[1]   1.8829655   0.6316624 111.0344483  40.4431900
0.8759374
>
> cor(x,y)*sd(y)/sd(x)
[1] 56.08328
>
> mean(y)-(cor(x,y)*sd(y)/sd(x))*mean(x)
[1] 5.431568
>

# Add the line to the plot
plot(x, y, xlab="square footage (1000 sq feet)",
ylab="price ($1000)")
abline(house.reg)
```

Fitted regression for house price data



The estimated regression coefficients are:

$$\hat{\beta}_0 = 5.432, \; \hat{\beta}_1 = 56.083$$

Q: How do we interpret these coefficients? What is the predicted price of a house that is 3200 square feet?

Issue: How well does the fitted regression line describe the data?

Approach 1: Consider r^2.

- High r^2 (and hence $|r|$) $\implies$ points are tightly clustered around the line $\implies$ predicted Ys are close to observed Ys $\implies$ residuals are small $\implies$ fit is good

Approach 2: Consider the variability in Ys explained by regression. To understand this, let's think about why the house prices are different. This is because the houses may have

- different square-footage
- different locations
- different years of sale
- other known/unknown reasons

Analysis of Variance (ANOVA)

- Total variability in Ys:
  $SS_{TOT} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = (n-1)S_y^2$ — **total SS**
- A part of SS_{TOT} is explained by the fitted regression:
  $SS_{REG} = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ — **SS due to regression**
- The rest is error variability:
  $SS_{ERR} = SS_{TOT} - SS_{REG} = \sum_{i=1}^{n} e_i^2$ — **error SS**
- **ANOVA Identity:** $SS_{TOT} = SS_{REG} + SS_{ERR}$.

This suggests *proportion of total variation explained*,

$$R^2 = \frac{SS_{REG}}{SS_{TOT}}$$

as a measure of **goodness of fit** of the fitted regression.

- Also called **coefficient of determination**
- Between 0 and 1, with high values suggesting a good fit.

Simple linear regression $(E(Y|x) = \beta_0 + \beta_1 x)$

- $\mathrm{SS}_{\mathrm{TOT}} = (n-1)S_y^2$
- $\mathrm{SS}_{\mathrm{REG}} = r^2(n-1)S_y^2$
- $\mathrm{SS}_{\mathrm{ERR}} = (1-r^2)(n-1)S_y^2$
- $R^2 = r^2$ — a reasonable measure from Approach 1 also.

Ex: For house price data: $r^2 = 0.88^2 \approx 0.77$

Alternative form for a regression model

Regression model: Models mean response — $E(Y|X = x)$ — as a function of x

Alternative form: $Y = E(Y|X = x) + \epsilon$

- $E(Y|x)$ is modeled as before
- $\epsilon = Y - E(Y|X = x) =$ **error** — a catchall for everything that causes the observed response to differ from its mean — e.g., random variability, effect of missing predictors, etc.
- $E(\epsilon) = 0$, $\text{var}(\epsilon) = \sigma^2$

Model for data: $Y_i = E(Y_i|X = x_i) + \epsilon_i$, $i = 1, \ldots, n$

Regression assumptions: The errors ϵ_i have mean zero, variance σ^2, and are independent. No additional assumptions are needed to estimate regression coefficients by least squares.

Additional assumption: Errors follow a **normal** distribution — needed for testing hypotheses and constructing confidence intervals. This means

$$\epsilon_i \sim \text{i.i.d. } N(0, \sigma^2), \ i = 1, \ldots, n$$

Simple Linear Regression with Normality

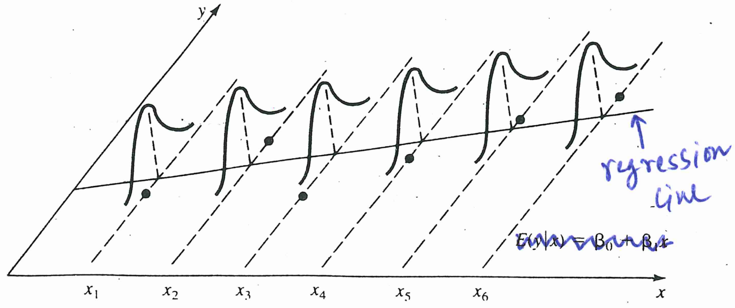
Assumed model: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $\epsilon_i \sim$ i.i.d. $N(0, \sigma^2)$, $i = 1, \ldots, n$.

Note: The values $x_1, \ldots, x_n$ of predictor X are known and fixed (i.e., non-random), and are assumed to be measured without error.

Properties:

- $E(Y_i | x_i) =$
- $\text{var}(Y_i | x_i) =$
- $Y_i | x_i \sim$ independent $N(\beta_0 + \beta_1 x_i, \sigma^2)$

Simple linear regression model



- The least squares estimators $(\hat{\beta}_0, \hat{\beta}_1)$ of (β_0, β_1) are also maximum likelihood estimators.
- $\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2/\{(n-1)S_x^2\}\right)$
-
- Define: $\hat{\sigma}^2 = SS_{\text{ERR}}/(n-2)$. Then, $E(\hat{\sigma}^2) = \sigma^2$.
- An unbiased estimator of σ^2 is
- **Note:** The sample variance S_y^2 is no longer unbiased for σ^2. This is because
- $SS_{\text{ERR}}/\sigma^2 = (n-2)\hat{\sigma}^2/\sigma^2$ follows a χ^2 distribution with $(n-2)$ degrees of freedom.

ANOVA table: A standard summary of regression fit. Here we have "simple linear regression" — i.e., two regression coefficients, β_0 and β_1.

Source	SS	d.f.	MS	F
Model	SS_{REG}	1	$MS_{\text{REG}} = \frac{SS_{\text{REG}}}{1}$	$\frac{MS_{\text{REG}}}{MS_{\text{ERR}}}$
Error	SS_{ERR}	$n-2$	$MS_{\text{ERR}} = \frac{SS_{\text{ERR}}}{n-2}$	
Total	SS_{TOT}	$n-1$		

Recall that:

- $\text{SS}_{\text{TOT}} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$
- $\text{SS}_{\text{REG}} = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$
- $\text{SS}_{\text{ERR}} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

Inference about slope β_1

Issue: Is the predictor X "significant", i.e., does it really help in predicting the response Y?

Approach 1: Test $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$. This is equivalent to testing $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$. (Why?)

Test statistic:

Null distribution:

- A two-sided t-test.

$100(1-\alpha)\%$ **Confidence Interval for β_1:**

Approach 2: Test for model significance. In simple linear regression, this is equivalent to testing $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.

Test statistic:

$$F = \frac{MS_{\mathrm{REG}}}{MS_{\mathrm{ERR}}}$$

Null distribution: This F statistic follows an F distribution with numerator d.f. 1 and denominator d.f. $n-2$.

- An F-test.
- Equivalent to the t-test seen before because $T^2 = F$ (verify).

Model evaluation

Issue: Is the fitted model a good representation of the data?

Approach: Examine the residuals, $e_i = Y_i - \hat{Y}_i$, $i = 1, \ldots, n$, and verify the key assumptions, namely,

- Errors have mean zero and constant variance
- Errors are normally distributed
- Errors are independent — often an issue when the data are collected over time.

Key Graphical Tools:

- **Residual plot:** Plot of residuals e_i against fitted values $\hat{Y}_i$. In the ideal plot, the points are scattered around zero and there is no pattern. This verifies the first assumption.
- **Normal QQ plot:** This verifies the normality assumption.
- **Time series plot:** Plot e_i against i. In the ideal plot, there should be no dependence, which verifies the independence assumption. More sophisticated tools exist.

Ex: House price data, continued.

```
x <- house$size
y <- house$price

house.reg <- lm (y ~ x)

# ANOVA table

> (anova(house.reg))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1  71534   71534  184.62 < 2.2e-16 ***
Residuals 56  21698     387
---
Signif. codes: 0‘***’0.001‘**’0.01‘*’0.05‘.’0.1‘ ’1
```

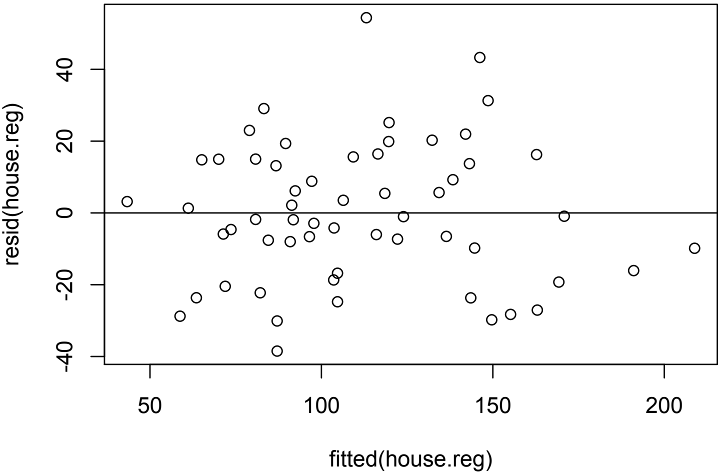
```
>

# Testing for zero slope

> summary(house.reg)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-38.489 -14.512  -1.422  14.919  54.389

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.432      8.191   0.663     0.51
x             56.083      4.128  13.587   <2e-16 ***
---
```

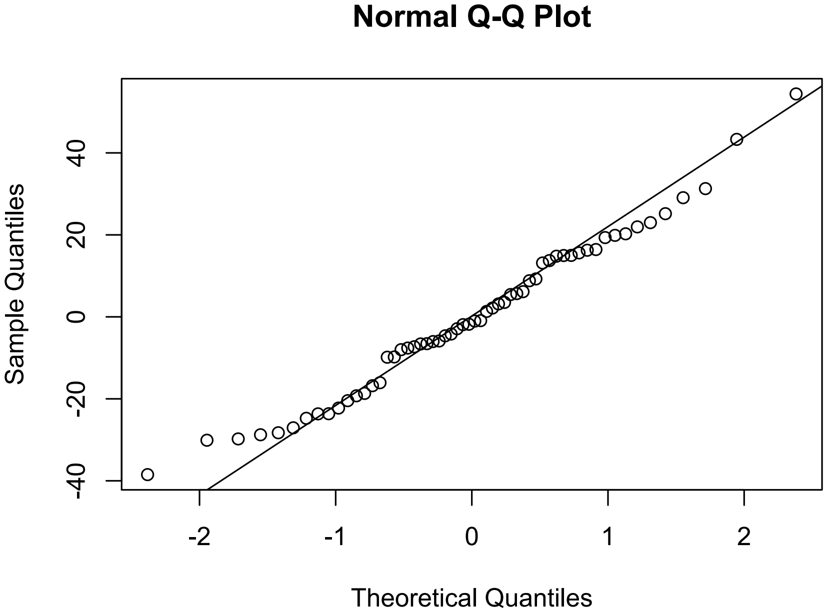
```
Signif. codes: 0‘***’0.001‘**’0.01‘*’0.05‘.’0.1‘ ’1

Residual standard error: 19.68 on 56 degrees of freedom
Multiple R-squared: 0.7673,Adjusted R-squared: 0.7631
F-statistic: 184.6 on 1 and 56 DF,  p-value: < 2.2e-16

>

# Confidence interval for slope

> confint(house.reg)
                2.5 %   97.5 %
(Intercept) -10.97619 21.83933
x            47.81473 64.35183
>

# Prediction at a new x
```

```
x.new <- data.frame(x=3)

> (predict(house.reg, newdata=x.new))
       1
173.6814
>


# Use fitted(house.reg) to get the fitted values
# Use resid(house.reg) to get the residuals


# Residual plot

plot(fitted(house.reg), resid(house.reg))
abline(h=0)

# QQ plot
```

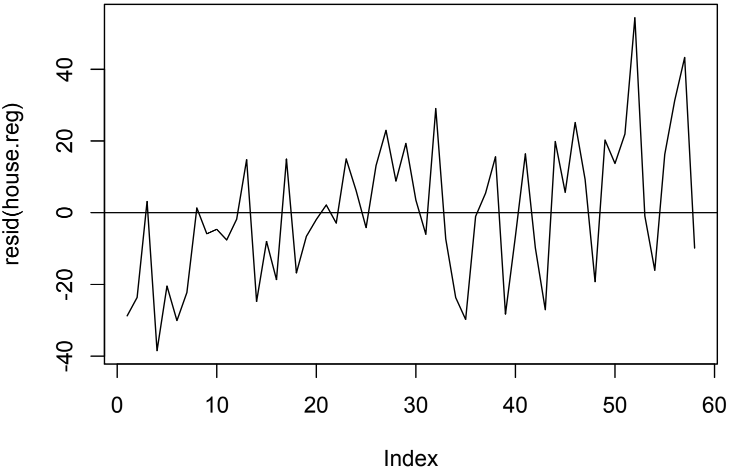
```
qqnorm(resid(house.reg))
qqline(resid(house.reg))


# Time series plot of residuals

plot(resid(house.reg), type="l")
abline(h=0)
```

Residual plot




Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-40
-20
0
20
40
-2
-1
0
1
2

Time series plot



Multiple Linear Regression

Simple linear regression: One predictor — X

Multiple linear regression: Several predictors — $X_1, \ldots, X_k$

Linear (regression) model:
$E(Y|X_1 = x_1, \ldots, X_k = x_k) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ — models **mean response** as a function of predictors

Examples:

- $E(Y|x) = \beta_0 + \beta_1 x$
- $E(Y|\mathbf{x}) = \beta_0 + \beta_1 x + \beta_2 x^2$
- $E(Y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2)$
- $E(\log(Y)|x) = \beta_0 + \beta_1 \log(x)$ —
- $E(Y|x) = \beta_0 + (\beta_1 x)^{-1}$ —

Note: "Linear" refers to **linear in regression coefficients**

Linear model: $E(Y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$

Interpretation of $k+1$ regression coefficients:

- $\beta_0 = E(Y|\mathbf{x} = \mathbf{0})$ — intercept
- $\beta_j = E(Y|x_1, \ldots, x_j + 1, \ldots, x_k) - E(Y|x_1, \ldots, x_j, \ldots, x_k)$ — slope of x_j, i.e., change in mean response when jth predictor increases by 1, while keeping other predictors fixed, $j = 1, \ldots, k$.

Data: n independent subjects, ith subject gives $(Y_i, X_{1i}, X_{2i}, \ldots, X_{ki})$, $i = 1, \ldots, n$.

Linear model for data: For $i = 1, \ldots, n$,
$E(Y_i|x_{i1}, \ldots, x_{ik}) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$

Alternative form: $Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i$

Assumptions:

- $E(\epsilon_i) = 0$, $\mathrm{var}(\epsilon_i) = \sigma^2$, and ϵ_i are independent.
- $k + 1 < n$ — i.e., have more observations than the number of regression coefficients
- The predictors are considered fixed and are measured without error

These imply:

- $E(Y_i | x_{i1}, \ldots, x_{ik}) =$
- $\mathrm{var}(Y_i) =$
- $Y_1, \ldots, Y_n$ are independent.

Linear Model in Matrix Notation

Define:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- $Y_i =$ BXi
- $\mathbf{Y} =$
- $E(\mathbf{Y}|\mathbf{X}) =$
- rank of X is full, i.e., $(\mathbf{X}'\mathbf{X})^{-1}$ exists.
- $\hat{\boldsymbol{\beta}}$ = estimator of $\boldsymbol{\beta}$
- $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ = fitted (or predicted) response

Predicted response when $\mathbf{x} = \mathbf{x}_0$: $\hat{Y}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}}$

Least Squares Estimation of $\boldsymbol{\beta}$

As before: Minimize $\sum_{i=1}^{n} \epsilon_i^2$ with respect to $\beta_0, \beta_1, \beta_k$ to get $\hat{\boldsymbol{\beta}}$

- Least squares estimator: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$
- Minimum value of $\sum_{i=1}^{n} \epsilon_i^2$ is $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = SS_{\text{ERR}}$ — **error (or residual) sum of squares**

Properties of $\hat{\boldsymbol{\beta}}$:

- Linear in $\mathbf{Y}$
- Unbiased, i.e.,
- $\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$
- $\text{var}(\hat{\beta}_0) = \sigma^2 \times$ first diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$
- $\text{var}(\hat{\beta}_j) = \sigma^2 \times (j+1)$th diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$
- $\hat{\sigma}^2 = SS_{\text{ERR}}/(n-k-1) = MS_{\text{ERR}}$ is unbiased for σ^2.

ANOVA table

As before:

- $\text{SS}_{\text{TOT}} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = (\mathbf{Y} - \overline{\mathbf{Y}})'(\mathbf{Y} - \overline{\mathbf{Y}})$, where
$$\overline{\mathbf{Y}} = \overline{Y}\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$
- $\text{SS}_{\text{REG}} = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = (\hat{\mathbf{Y}} - \overline{\mathbf{Y}})'(\hat{\mathbf{Y}} - \overline{\mathbf{Y}})$
- $\text{SS}_{\text{ERR}} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}})$

Source	SS	d.f.	MS	F
Model	SS_{REG}	k	$MS_{\text{REG}} = \frac{SS_{\text{REG}}}{k}$	$\frac{MS_{\text{REG}}}{MS_{\text{ERR}}}$
Error	SS_{ERR}	$n-k-1$	$MS_{\text{ERR}} = \frac{SS_{\text{ERR}}}{n-k-1}$	
Total	SS_{TOT}	$n-1$		

Testing hypotheses: Additionally assume that ϵ_i are normal, i.e., $\epsilon_i \sim$ independent $N(0, \sigma^2)$.

- Each $\hat{\beta}_j$ follows a normal distribution
- $(n-k-1)\hat{\sigma}^2/\sigma^2$ follows a χ^2_{n-k-1} distribution

Testing significance of jth predictor:

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0$$

- H_0 : Predictor X_j is not useful for predicting response **after adjusting for the other predictors**
- Test statistic:
- Null distribution:
- Rejection region:
- p-value:
- $100(1-\alpha)\%$ CI for β_j:

Testing model significance:

$$H_0 : \beta_1 = \ldots = \beta_k = 0 \;\; \text{vs.} \;\; H_1 : \text{at least one } \beta_j \neq 0$$

- H_0 : None of the predictors is useful for predicting response
- Test statistic:
- Null distribution:
- Rejection region:
- p-value:

Coefficient of determination: As before,

$$R^2 = \frac{SS_{\text{REG}}}{SS_{\text{TOT}}} = 1 - \frac{SS_{\text{ERR}}}{SS_{\text{TOT}}}$$

- R^2 = proportion of total variation explained by regression
- R^2 = square of correlation between $(Y_i, \hat{Y}_i)$, $i = 1, \ldots, n$ — can be verified
- When we add new predictors to the model, R^2
- Even adding useless predictors will
- Not a fair criterion for comparing models with different numbers of predictors

Beware of overfitting a model:

- Overfitting = Having too many predictors in the model
- An overfitted model will provide a good fit to the data at hand, but it will be terrible at predicting future observations.
- Estimated regression coefficients have large standard errors.
- See Figure 11.4 on page 369

Adjusted R^2:

$$R^2_{\text{adj}} = 1 - \frac{SS_{\text{ERR}}/(n-k-1)}{SS_{\text{TOT}}/(n-1)}$$

- Unlike R^2, R^2_{adj} rewards adding a predictor only if it reduces the error SS considerably
- Imagine adding a useless predictor. In this case, SS_{REG} and hence SS_{ERR} does not change. However, $SS_{\text{ERR}}/(n-k-1)$ increases, which in turn decreases R^2_{adj}
- A more fair measure of goodness-of-it than R^2
- Can be used to compare two models with different numbers of predictors — choose the model with the highest R^2_{adj}

Comparing two Nested Models

Nested models: Model 2 is nested within Model 1 if the predictors of Model 2 are a subset of predictors of Model 1.

Issue: How to compare two nested models?

Full model: Predictors $X_1, \ldots, X_m$

Reduced model: Predictors $X_1, \ldots, X_k$, i.e., it does not have predictors $X_{k+1}, \ldots, X_m$

Hypotheses: $H_0 : \beta_{k+1} = \ldots = \beta_m = 0$, vs., H_1 : at least one slope $\neq 0$

Extra sum of squares: Difference in variation explained by the two models

$$\begin{aligned} SS_{\text{EX}} &= SS_{\text{REG}}(\text{full}) - SS_{\text{REG}}(\text{reduced}) \\ &= SS_{\text{ERR}}(\text{reduced}) - SS_{\text{ERR}}(\text{full}) \end{aligned}$$

- SS_{EX} has $m-k$ degrees of freedom. It equals the number of regression coefficients set to zero under H_0.

Test statistic:

$$F = \frac{MS_{EX}}{MS_{\text{ERR}}(\text{full})} = \frac{SS_{EX}/(m-k)}{SS_{\text{ERR}}(\text{full})/(n-m-1)}$$

Null distribution:

Rejection region:

p**-value:**

- aka "partial F-test"
- Used for designing stepwise model selection procedures (see pages 392-394)

Example: Home price data. These data come from a sample of homes sold in Maplewood, NJ in 2001.

```
# Read the home price data

home <- read.table("homeprice_multiple_predictors.txt",
sep=",", header=T)

> str(home)
'data.frame': 29 obs. of  7 variables:
 $ list        : num  80 151 310 295 339 ...
 $ sale        : num  118 151 300 275 340 ...
 $ full        : int  1 1 2 2 2 1 3 1 1 1 ...
 $ half        : int  0 0 1 1 0 1 0 1 2 0 ...
 $ bedrooms    : int  3 4 4 4 3 4 3 3 3 1 ...
 $ rooms       : int  6 7 9 8 7 8 7 7 7 3 ...
 $ neighborhood: int  1 1 3 3 4 3 2 2 3 2 ...
>
```

```
# Attach the dataset in R's memory so that we can
# directly use the names of the variables

attach(home)

# Look at distributions of some predictors

> table(bedrooms)
bedrooms
 1  2  3  4  5
 1  3 16  8  1
>
> table(full)
full
 1  2  3
13 11  5
>
```

```
> table(half)
half
 0  1  2
13 13  3
>
> table(neighborhood)
neighborhood
 1  2  3  4  5
 2  8 12  5  2
>

# Regress sale price on # bedrooms and neighborhood

fit1 <- lm(sale ~ bedrooms + neighborhood)

> summary(fit1)

Call:
```

```
lm(formula = sale ~ bedrooms + neighborhood)

Residuals:
    Min      1Q  Median      3Q     Max 
-90.871 -39.861   0.636  28.815 107.660 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -132.057     40.341  -3.273 0.003001 ** 
bedrooms       42.483     11.446   3.712 0.000987 ***
neighborhood   93.493      9.101  10.273 1.21e-10 ***
---
Signif. codes: 0‘***’0.001‘**’0.01‘*’0.05‘.’0.1‘ ’1

Residual standard error: 47.3 on 26 degrees of freedom
Multiple R-squared: 0.8491,Adjusted R-squared: 0.8375
F-statistic: 73.16 on 2 and 26 DF,  p-value: 2.1e-11
>
```

```
# Add # full and half baths

fit2 <- update(fit1, . ~ . + full + half)

> summary(fit2)

Call:
lm(formula = sale ~ bedrooms + neighborhood + full + half)

Residuals:
    Min      1Q  Median      3Q     Max
-56.554 -38.067   6.027  26.998  53.311

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -125.121     33.136  -3.776 0.000926 ***
bedrooms      29.513     10.091   2.925 0.007419 **
```

```
neighborhood   78.724      9.669   8.142 2.31e-08 ***
full           27.345     13.604   2.010 0.055785 .
half           45.553     12.129   3.756 0.000974 ***
---
Signif. codes: 0‘***’0.001‘**’0.01‘*’0.05‘.’0.1‘ ’1

Residual standard error: 38.79 on 24 degrees of freedom
Multiple R-squared: 0.9063,Adjusted R-squared: 0.8907
F-statistic: 58.05 on 4 and 24 DF,  p-value: 5.425e-12
>

# Drop # full baths

fit3 <- update(fit2, . ~ . - full)

> summary(fit3)

Call:
```

```
lm(formula = sale ~ bedrooms + neighborhood + half)

Residuals:
   Min     1Q Median     3Q    Max
-67.55 -42.27   7.17  26.93  68.83

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -127.348     35.073  -3.631  0.00127 **
bedrooms       35.649     10.187   3.500  0.00177 **
neighborhood   90.982      7.947  11.449 1.95e-11 ***
half           37.004     12.030   3.076  0.00503 **
---
Signif. codes: 0'***'0.001'**'0.01'*'0.05'.'0.1' '1

Residual standard error: 41.08 on 25 degrees of freedom
Multiple R-squared: 0.8905,Adjusted R-squared: 0.8774
F-statistic:  67.8 on 3 and 25 DF,  p-value: 3.808e-12
```

```
>

# Compare the nested models

Check {\tt ?anova.lm}
```

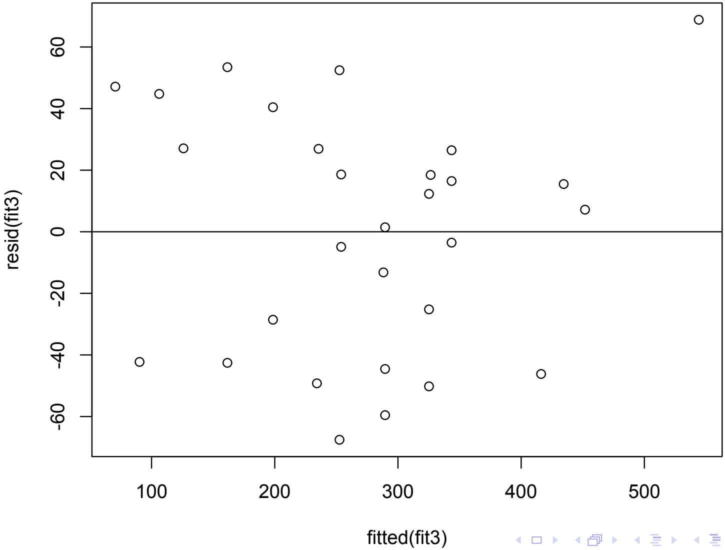
Important note: When comparing two models using `anova` the results are as expected from the partial F-test. However, when more than two models are compared using `anova`, the F-statistic and p-value may not be what we would like. The reason for this is that the F-statistic compares the mean SS for a row to the MS_{ERR} for the **largest model** considered.

```
> anova(fit1, fit3, fit2)
Analysis of Variance Table

Model 1: sale ~ bedrooms + neighborhood
Model 2: sale ~ bedrooms + neighborhood + half
Model 3: sale ~ bedrooms + neighborhood + full + half
  Res.Df   RSS Df Sum of Sq       F   Pr(>F)
1     26 58164
2     25 42194  1   15970.1 10.6132 0.003338 **
3     24 36114  1    6080.1  4.0406 0.055785 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
>

> anova(fit1, fit2)
Analysis of Variance Table

Model 1: sale ~ bedrooms + neighborhood
```

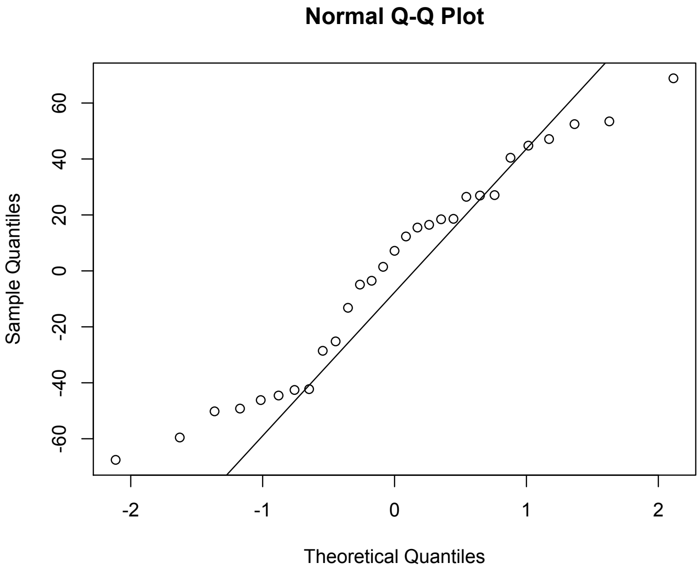
```
Model 2: sale ~ bedrooms + neighborhood + full + half
  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1     26 58164
2     24 36114  2     22050 7.3269 0.003283 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
>

> anova(fit3, fit2)
Analysis of Variance Table

Model 1: sale ~ bedrooms + neighborhood + half
Model 2: sale ~ bedrooms + neighborhood + full + half
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     25 42194
2     24 36114  1    6080.1 4.0406 0.05579 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
```

```
>

# Residual plot

plot(fitted(fit3), resid(fit3))
abline(h=0)

# QQ plot

qqnorm(resid(fit3))
qqline(resid(fit3))
```

Residual plot

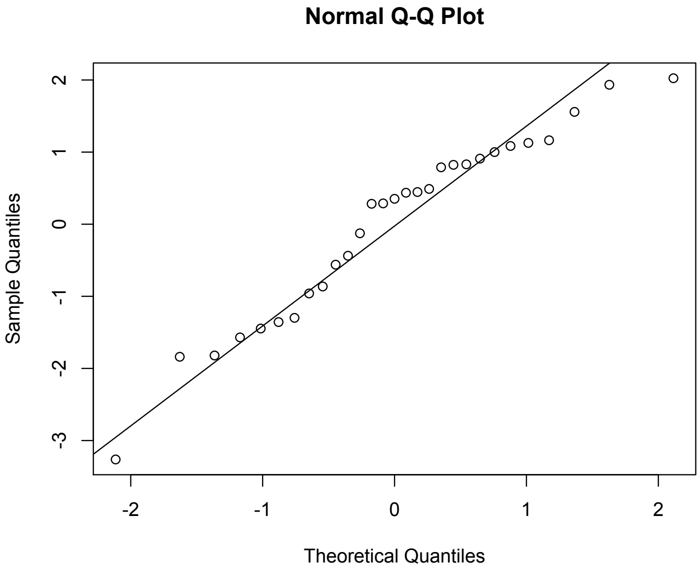


Normal QQ plot



```r
# Take sqrt(sale) rather than sale as response

fit4 <- update(fit3, sqrt(sale) ~ .)

# New QQ plot

qqnorm(resid(fit4))
qqline(resid(fit4))
```

Normal QQ plot for transformed data



Regression with categorical predictors

Categorical variable: Its values are categories (or attributes) with no particular order, e.g., race, OS, etc. The values should **not** be coded as 1, 2, 3, .., unless `R` knows to treat the variable as a `factor`.

Dummy variable: A binary variable z with value 0 or 1

Key idea: Represent a categorical variable with C categories using $C - 1$ dummy variables, $z_1, \ldots, z_{C-1}$. The model is

$$E(Y|\mathbf{z}) = \beta_0 + \gamma_1 z_1 + \ldots + \gamma_{C-1} z_{C-1}.$$

Base (or reference) category: $z_1 = \ldots = z_{C-1} = 0$.

Ex 1: OS with two categories — Windows and Mac.

Ex 2: Race with three categories — White, Black and other.

- In general, β_0 = mean for base category, and β_j = difference in means for category j and base category
- The regression model may have both numerical as well as categorical predictors.
- The model may have several categorical predictors.
- To test whether a categorical variable is significant, simultaneously test **all** corresponding slopes. In other words, the hypotheses are $H_0 : \gamma_1 = \ldots = \gamma_{C-1} = 0$, vs. H_1 : at least one non-zero slope, and they should be tested using an F-test with $C - 1$ numerator d.f.

Example: Jane data.

```
# Read the Jane data

jane <- read.table("jane.csv", sep=",", header=T)

> str(jane)
'data.frame': 150 obs. of  3 variables:
 $ x    : int  1 1 1 2 2 2 3 3 3 4 ...
 $ color: Factor w/ 3 levels "blue","green",..: 3 1 2 3 1 2
 $ y    : num  24.9 12.3 16.6 25.2 12.1 ...
>

attach(jane)

> table(color)
color
 blue green   red
```

```
    50     50     50
>

# Include both x and color as predictors

fit1 <- lm(y~x+color)

# Note: color is already a factor variable. If this is
# numeric, then we need to write:
# fit1 <- lm(y~ x + factor(color))

> summary(fit1)

Call:
lm(formula = y ~ x + color)

Residuals:
     Min        1Q    Median        3Q       Max
```

```
-14.2398  -2.9939   0.1725   3.5555  11.9747

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.16989    1.01710  12.948  < 2e-16 ***
x            1.00344    0.02848  35.227  < 2e-16 ***
colorgreen   2.12586    1.00688   2.111   0.0364 *
colorred     6.60586    1.00688   6.561  8.7e-10 ***
---
Signif. codes: 0‘***’0.001‘**’0.01‘*’0.05‘.’0.1‘ ’1

Residual standard error: 5.034 on 146 degrees of freedom
Multiple R-squared: 0.898,Adjusted R-squared: 0.8959
F-statistic: 428.6 on 3 and 146 DF,  p-value: < 2.2e-16
>

# Is color significant?
```

```
fit2 <- lm(y~x)

> anova(fit2, fit1)
Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ x + color
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    148 4837.5
2    146 3700.4  2    1137.1 22.433 3.197e-09 ***
---
Signif. codes: 0'***'0.001'**'0.01'*'0.05'.'0.1' '1
>
```

Q: What is the predicted response for a subject with color=blue and x=2?