MATH 4070 R Session 2: Multiple Linear Regression I

Whitney

Contents

Species Diversity on the Galápagos Islands: Data Exploration 2
- First Step: Load the data 2
- Plot the pairwise scatterplots 2
- Correlation matrix 3
- Use *ggpairs* for scatterplots and correlation 3

Fitting Linear Regression Models 4
- Model 1: Fitting a simple linear regression 4
- Model 2: Adding *Area* 6
- Model 3: Adding *Adjacent* 8
- Full Model 9
- Parameter Estimation 10

Regression with Both Numerical and Categorical Predictors 11
- Salaries for Professors Data Set 11
- Load the data 11
- Model Fitting 11

Polynomial regression 15
- Housing Values in Suburbs of Boston 15
- Load and plot the data 15
- Plot the poylnomial regression fits 17

ANOVA 19

Simulation 20
- Step 1: Simulate the data sets 20
- Step 2: Compute R^2 and R^2_{adj} for Model 1 and Model 2 20
- Compare R^2 for for Model 1 and Model 2 20
- Compare R^2_{adj} for for Model 1 and Model 2 21

Species Diversity on the Galápagos Islands: Data Exploration

First Step: Load the data

You will need to install the R package faraway using `install.packages("faraway")`. This only needs to be done once. After that, load the package with `library(faraway)`, which must be done every time you use it.

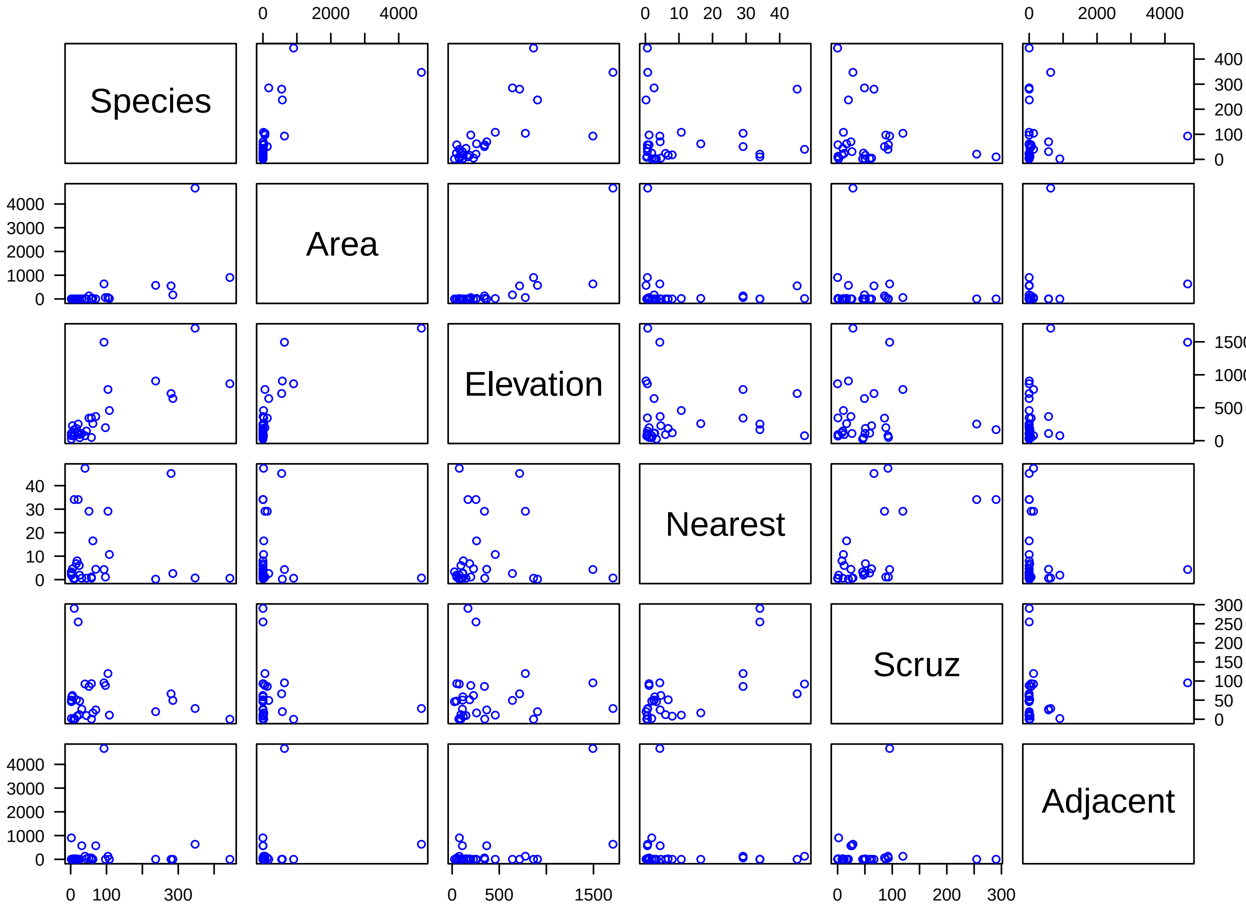
```r
#install.packages("faraway")
library(faraway)
data(gala)
head(gala)
```

```
##              Species Endemics  Area Elevation Nearest Scruz Adjacent
## Baltra            58       23 25.09       346     0.6   0.6     1.84
## Bartolome         31       21  1.24       109     0.6  26.3   572.33
## Caldwell           3        3  0.21       114     2.8  58.7     0.78
## Champion          25        9  0.10        46     1.9  47.4     0.18
## Coamano            2        1  0.05        77     1.9   1.9   903.82
## Daphne.Major      18       11  0.34       119     8.0   8.0     1.84
```

For the remaining analysis, we will remove the variable *Endemics* as it is highly correlated with our response variable, *Species*.

Plot the pairwise scatterplots

```r
pairs(gala[, -2], cex = 0.95, col = "blue", las = 1)
```



Correlation matrix

```
cor(gala[, -2])
```

```
##                Species       Area   Elevation     Nearest       Scruz
## Species     1.00000000  0.6178431  0.73848666 -0.01409407 -0.17114244
## Area        0.61784307  1.0000000  0.75373492 -0.11110320 -0.10078493
## Elevation   0.73848666  0.7537349  1.00000000 -0.01107698 -0.01543829
## Nearest    -0.01409407 -0.1111032 -0.01107698  1.00000000  0.61541036
## Scruz      -0.17114244 -0.1007849 -0.01543829  0.61541036  1.00000000
## Adjacent    0.02616635  0.1800376  0.53645782 -0.11624788  0.05166066
##               Adjacent
## Species     0.02616635
## Area        0.18003759
## Elevation   0.53645782
## Nearest    -0.11624788
## Scruz       0.05166066
## Adjacent    1.00000000
```

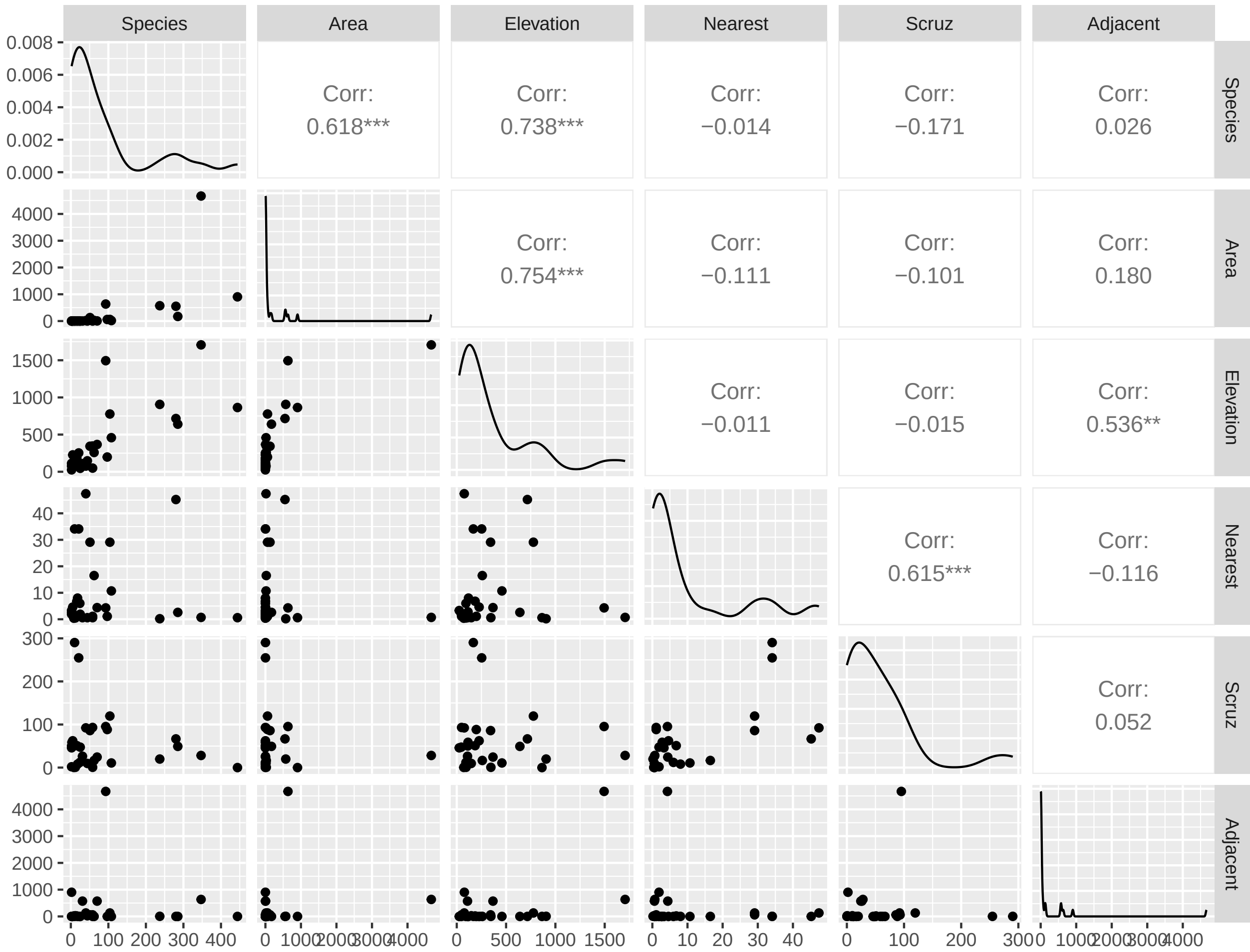
Use *ggpairs* for scatterplots and correlation

Scatterplots of each pair are visualized in the lower-left panels, while Pearson correlation values and significance are displayed in the upper-right panels.

```r
library(GGally)
```

```
## Loading required package: ggplot2

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

##
## Attaching package: 'GGally'

## The following object is masked from 'package:faraway':
##
##     happy
```

```r
ggpairs(gala[, -2])
```



Fitting Linear Regression Models

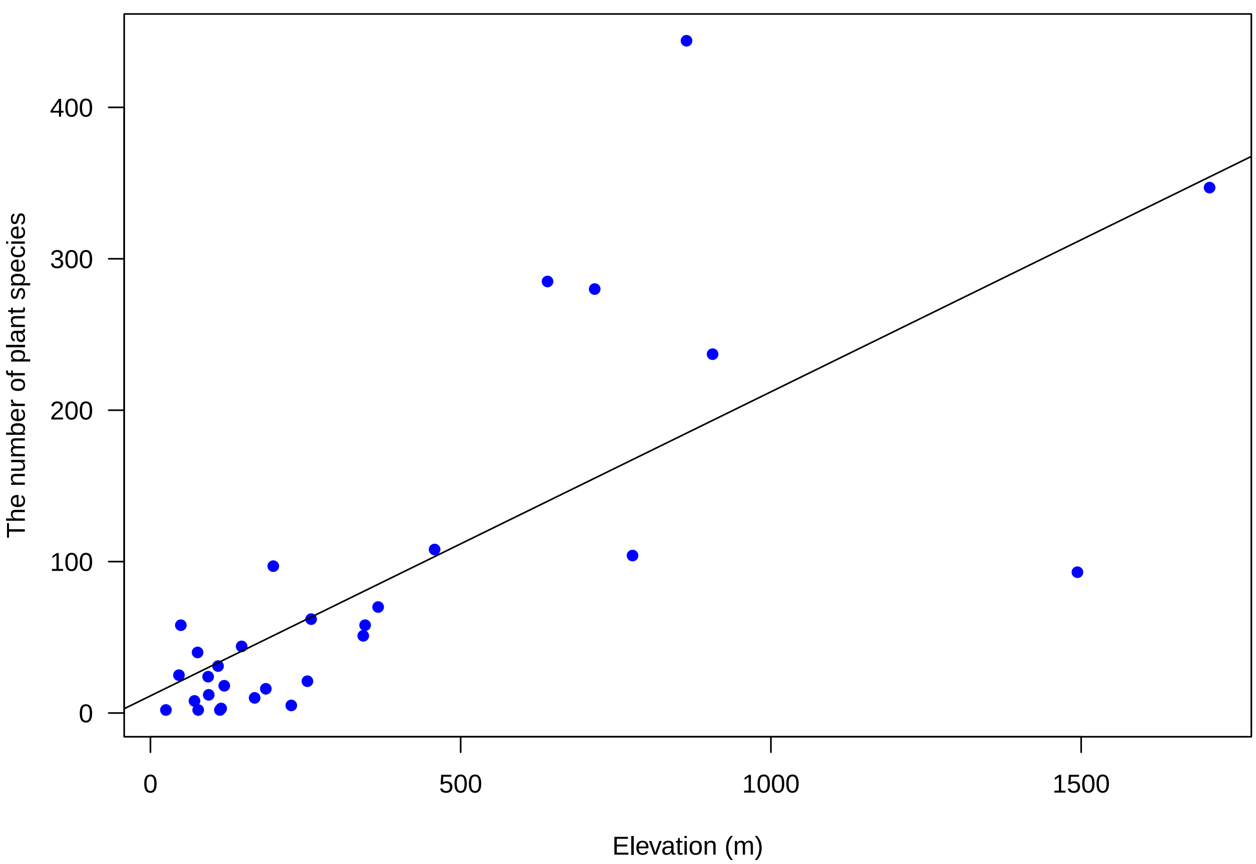
Model 1: Fitting a simple linear regression

Here we use *Elevation* as the predictor as it has the highest correlation with *Species*

```r
M1 <- lm(Species ~ Elevation, data = gala)
summary(M1)
```

```
##
## Call:
## lm(formula = Species ~ Elevation, data = gala)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -218.319  -30.721  -14.690    4.634  259.180
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.33511   19.20529   0.590     0.56
## Elevation    0.20079    0.03465   5.795 3.18e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78.66 on 28 degrees of freedom
## Multiple R-squared:  0.5454, Adjusted R-squared:  0.5291
## F-statistic: 33.59 on 1 and 28 DF,  p-value: 3.177e-06
```

```r
plot(gala$Elevation, gala$Species, xlab = "Elevation (m)",
     ylab = "The number of plant species",
     las = 1, pch = 16, col = "blue")
abline(M1)
```

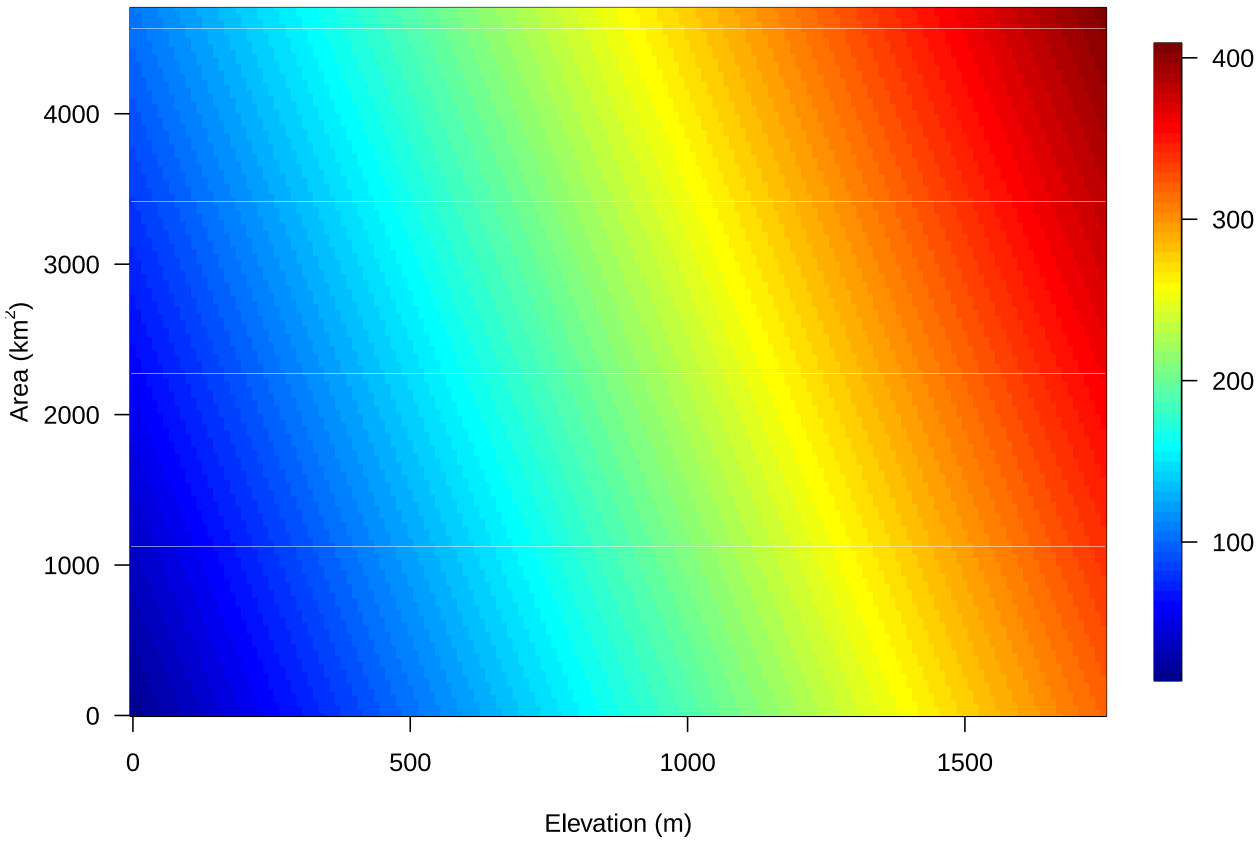


Model 2: Adding *Area*

```r
M2 <- lm(Species ~ Elevation + Area, data = gala)
summary(M2)
```

```
##
## Call:
## lm(formula = Species ~ Elevation + Area, data = gala)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -192.619  -33.534  -19.199    7.541  261.514
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.10519   20.94211   0.817  0.42120
## Elevation    0.17174    0.05317   3.230  0.00325 **
## Area         0.01880    0.02594   0.725  0.47478
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 79.34 on 27 degrees of freedom
## Multiple R-squared:  0.554,  Adjusted R-squared:  0.521
## F-statistic: 16.77 on 2 and 27 DF,  p-value: 1.843e-05
```

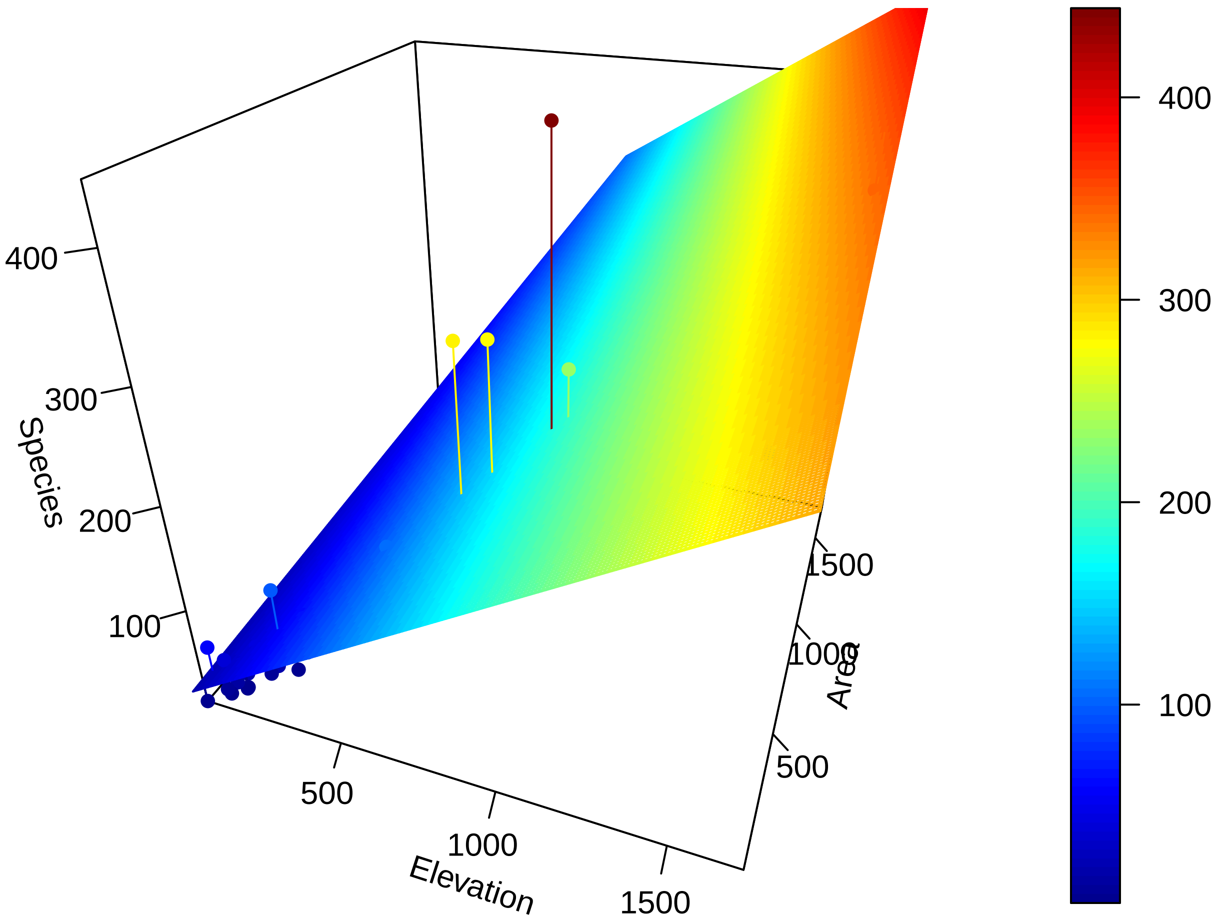
```r
library(fields)
Elevation_grid <- seq(0, 1750, 10)
Area_grid <- seq(0, 4700, 10)
temp <- expand.grid(Elevation_grid, Area_grid)
x_new <- data.frame(Elevation = temp$Var1, Area = temp$Var2)

y_pred <- matrix(predict(M2, x_new), nrow = length(Elevation_grid))
image.plot(Elevation_grid, Area_grid,  y_pred, las = 1,
           xlab = "Elevation (m)", ylab = expression(paste("Area (", km^2, ")")))
```




```r
library(plot3D)
```

```
## Warning in fun(libname, pkgname): couldn't connect to display
## "/private/tmp/com.apple.launchd.YHAOSUBV6c/org.xquartz:0"
```

```r
# fitted points for droplines to surface
fitpoints <- predict(M2)
# scatter plot with regression plane
scatter3D(gala$Elevation, gala$Elevation, gala$Species, pch = 16, cex = 1,
    theta = 20, phi = 30, ticktype = "detailed",
    xlab = "Elevation", ylab = "Area", zlab = "Species",
    surf = list(x = Elevation_grid, y = Area_grid, z = y_pred, facets = NA, fit = fitpoints))
```



Model 3: Adding *Adjacent*

```
M3 <- lm(Species ~ Elevation + Area + Adjacent, data = gala)
summary(M3)
```

```
##
## Call:
## lm(formula = Species ~ Elevation + Area + Adjacent, data = gala)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -124.064  -34.283   -8.733   27.972  195.973
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.71893   16.90706  -0.338  0.73789
## Elevation    0.31498    0.05211   6.044  2.2e-06 ***
## Area        -0.02031    0.02181  -0.931  0.36034
## Adjacent    -0.07528    0.01698  -4.434  0.00015 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 61.01 on 26 degrees of freedom
## Multiple R-squared:  0.746,  Adjusted R-squared:  0.7167
## F-statistic: 25.46 on 3 and 26 DF,  p-value: 6.683e-08
```

Full Model

```r
M4 <- lm(Species ~ Elevation + Area + Adjacent + Nearest + Scruz, data = gala)
summary(M4)
```

```
##
## Call:
## lm(formula = Species ~ Elevation + Area + Adjacent + Nearest +
##     Scruz, data = gala)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -111.679  -34.898   -7.862   33.460  182.584
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.068221  19.154198   0.369 0.715351
## Elevation    0.319465   0.053663   5.953 3.82e-06 ***
## Area        -0.023938   0.022422  -1.068 0.296318
## Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
## Nearest      0.009144   1.054136   0.009 0.993151
## Scruz       -0.240524   0.215402  -1.117 0.275208
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.98 on 24 degrees of freedom
## Multiple R-squared:  0.7658, Adjusted R-squared:  0.7171
## F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

```r
predict(M4)
```

```
##       Baltra    Bartolome     Caldwell     Champion      Coamano Daphne.Major
##  116.7259460   -7.2731544   29.3306594   10.3642660  -36.3839155   43.0877052
## Daphne.Minor       Darwin         Eden     Enderby     Espanola   Fernandina
##   33.9196678   -9.0189919   28.3142017   30.7859425   47.6564865   96.9895982
##     Gardner1     Gardner2     Genovesa      Isabela     Marchena       Onslow
##   -4.0332759   64.6337956   -0.4971756  386.4035578   88.6945404    4.0372328
##        Pinta       Pinzon   Las.Plazas       Rabida SanCristobal  SanSalvador
##  215.6794862  150.4753750   35.0758066   75.5531221  206.9518779  277.6763183
##    SantaCruz      SantaFe   SantaMaria      Seymour      Tortuga         Wolf
##  261.4164131   85.3764857  195.6166286   49.8050946   52.9357316   26.7005735
```

```r
confint(M4)
```

```
##                   2.5 %      97.5 %
## (Intercept) -32.4641006 46.60054205
```

```
## Elevation     0.2087102  0.43021935
## Area         -0.0702158  0.02233912
## Adjacent     -0.1113362 -0.03827344
## Nearest      -2.1664857  2.18477363
## Scruz        -0.6850926  0.20404416
```

Parameter Estimation

```r
X <- model.matrix(M4)
y <- gala$Species
# regression parameters
(beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y)
```

```
##                     [,1]
## (Intercept)  7.068220709
## Elevation    0.319464761
## Area        -0.023938338
## Adjacent    -0.074804832
## Nearest      0.009143961
## Scruz       -0.240524230
```

```r
beta_hat_faster <- solve(crossprod(X), crossprod(X, y))
# fitted values
(y_hat <- X %*% solve(t(X) %*% X) %*% t(X) %*% y)
```

```
##                     [,1]
## Baltra       116.7259460
## Bartolome     -7.2731544
## Caldwell      29.3306594
## Champion      10.3642660
## Coamano      -36.3839155
## Daphne.Major  43.0877052
## Daphne.Minor  33.9196678
## Darwin        -9.0189919
## Eden          28.3142017
## Enderby       30.7859425
## Espanola      47.6564865
## Fernandina    96.9895982
## Gardner1      -4.0332759
## Gardner2      64.6337956
## Genovesa      -0.4971756
## Isabela      386.4035578
## Marchena      88.6945404
## Onslow         4.0372328
## Pinta        215.6794862
## Pinzon       150.4753750
## Las.Plazas    35.0758066
## Rabida        75.5531221
## SanCristobal 206.9518779
## SanSalvador  277.6763183
## SantaCruz    261.4164131
```

```
## SantaFe       85.3764857
## SantaMaria   195.6166286
## Seymour       49.8050946
## Tortuga       52.9357316
## Wolf          26.7005735
```

Regression with Both Numerical and Categorical Predictors

Salaries for Professors Data Set

The 2008-09 nine-month academic salary for Assistant Professors, Associate Professors and Professors in a college in the U.S. The data were collected as part of the on-going effort of the college's administration to monitor salary differences between male and female faculty members.

Load the data

```r
library(carData)
data(Salaries)
head(Salaries)
```

```
##        rank discipline yrs.since.phd yrs.service  sex salary
## 1      Prof          B            19          18 Male 139750
## 2      Prof          B            20          16 Male 173200
## 3  AsstProf          B             4           3 Male  79750
## 4      Prof          B            45          39 Male 115000
## 5      Prof          B            40          41 Male 141500
## 6 AssocProf          B             6           6 Male  97000
```

Model Fitting

```r
m1 <- lm(salary ~ discipline + rank + sex + yrs.since.phd, data = Salaries)
X <- model.matrix(m1)
head(X)
```

Model 1: A MLR with `yrs.since.phd` (numerical predictor), discipline, rank, and sex (categorical predictors)

```
##   (Intercept) disciplineB rankAssocProf rankProf sexMale yrs.since.phd
## 1           1           1             0        1       1            19
## 2           1           1             0        1       1            20
## 3           1           1             0        0       1             4
## 4           1           1             0        1       1            45
## 5           1           1             0        1       1            40
## 6           1           1             1        0       1             6
```

```r
summary(m1)
```

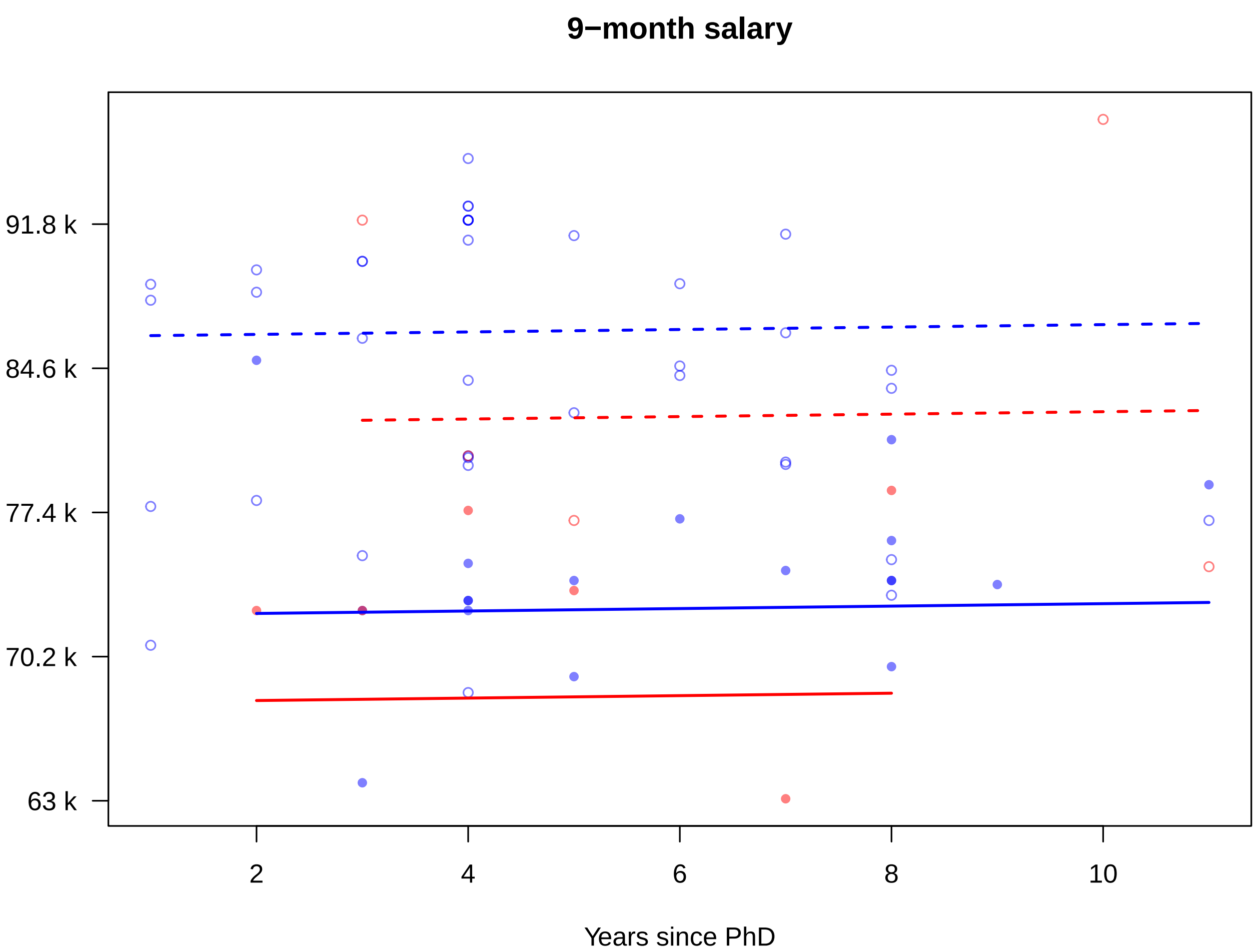
```
##
## Call:
## lm(formula = salary ~ discipline + rank + sex + yrs.since.phd,
##     data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -67451 -13860  -1549  10716  97023
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   67884.32    4536.89  14.963  < 2e-16 ***
## disciplineB   13937.47    2346.53   5.940 6.32e-09 ***
## rankAssocProf 13104.15    4167.31   3.145  0.00179 **
## rankProf      46032.55    4240.12  10.856  < 2e-16 ***
## sexMale        4349.37    3875.39   1.122  0.26242
## yrs.since.phd    61.01     127.01   0.480  0.63124
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22660 on 391 degrees of freedom
## Multiple R-squared:  0.4472, Adjusted R-squared:  0.4401
## F-statistic: 63.27 on 5 and 391 DF,  p-value: < 2.2e-16
```

```r
attach(Salaries)
yr.range <- tapply(yrs.since.phd, list(discipline, sex, rank), range)
sex.col <- ifelse(sex == "Male", "blue", "red")
dis.col <- ifelse(discipline == "A", 16, 1)

beta0 <- m1$coefficients[1]
betaDisp <- m1$coefficients[2]
betaAssoc <- m1$coefficients[3]
betaProf <- m1$coefficients[4]
betaMale <- m1$coefficients[5]
beta1 <- m1$coefficients[6]
```

```r
library(scales)
# Plot the model fits by rank
## Assist prof
assistant <- which(rank == "AsstProf")
plot(yrs.since.phd[assistant], salary[assistant], pch = dis.col[assistant], cex = 0.8,
     col = alpha(sex.col[assistant], 0.5), yaxt = "n", xlab = "Years since PhD",
     main = "9-month salary", ylab = "")
axis(2, at = seq(63000, 99000, len = 6), labels = paste(seq(63000, 99000, len = 6)/ 1000, "k"),
     las = 1)

segments(yr.range[[1]][1], beta0 + yr.range[[1]][1] * beta1,
         yr.range[[1]][2], beta0 + yr.range[[1]][2] * beta1, col = "red", lwd = 1.8)
segments(yr.range[[2]][1], beta0 + betaDisp + yr.range[[2]][1] * beta1,
```

```r
         yr.range[[2]][2], beta0 + betaDisp + yr.range[[2]][2] * beta1,
         col = "red", lty = 2, lwd = 1.8)
segments(yr.range[[3]][1], beta0 + betaMale + yr.range[[3]][1] * beta1,
         yr.range[[3]][2], beta0 + betaMale + yr.range[[3]][2] * beta1,
         col = "blue", lwd = 1.8)
segments(yr.range[[4]][1], beta0 + betaDisp + betaMale + yr.range[[4]][1] * beta1,
         yr.range[[4]][2], beta0 + betaDisp + betaMale + yr.range[[4]][2] * beta1,
         col = "blue", lty = 2, lwd = 1.8)
```

Plot the Model 1 Fits




```r
m2 <- lm(salary ~ sex * yrs.since.phd)
summary(m2)
```

Model 2: Another MLR where we include the *interaction* between sex and `yrs.since.phd`

```
##
## Call:
## lm(formula = salary ~ sex * yrs.since.phd)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -83012 -19442  -2988  15059 102652
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            73840.8     8696.7   8.491 4.27e-16 ***
## sexMale                20209.6     9179.2   2.202 0.028269 *
## yrs.since.phd           1644.9      454.6   3.618 0.000335 ***
## sexMale:yrs.since.phd   -728.0      468.0  -1.555 0.120665
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27420 on 393 degrees of freedom
## Multiple R-squared:  0.1867, Adjusted R-squared:  0.1805
## F-statistic: 30.07 on 3 and 393 DF,  p-value: < 2.2e-16
```

```r
coeff <- m2$coefficients
plot(yrs.since.phd, salary, las = 1, pch = 16, cex = 0.5, col = alpha(sex.col, 0.5),
     xlab = "Years since PhD", main = "9-month salary", ylab = "")
grid()
abline(coeff[1], coeff[3], col = "red")
abline(coeff[1] + coeff[2], coeff[3] + coeff[4], col = "blue")
legend("toplef", legend = c("Female", "Male"),
       pch = 16, col = c("red", "blue"), bty = "n")
```

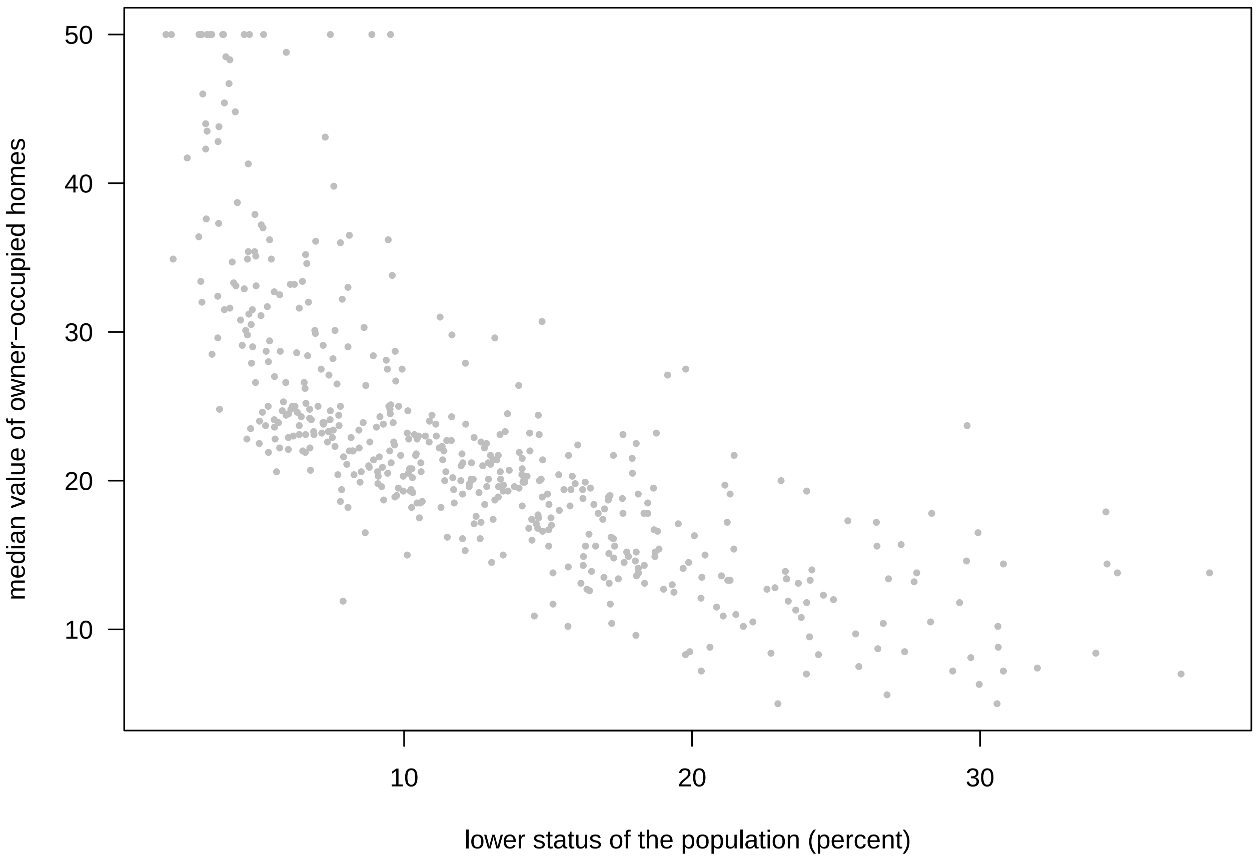
Polynomial regression

Housing Values in Suburbs of Boston

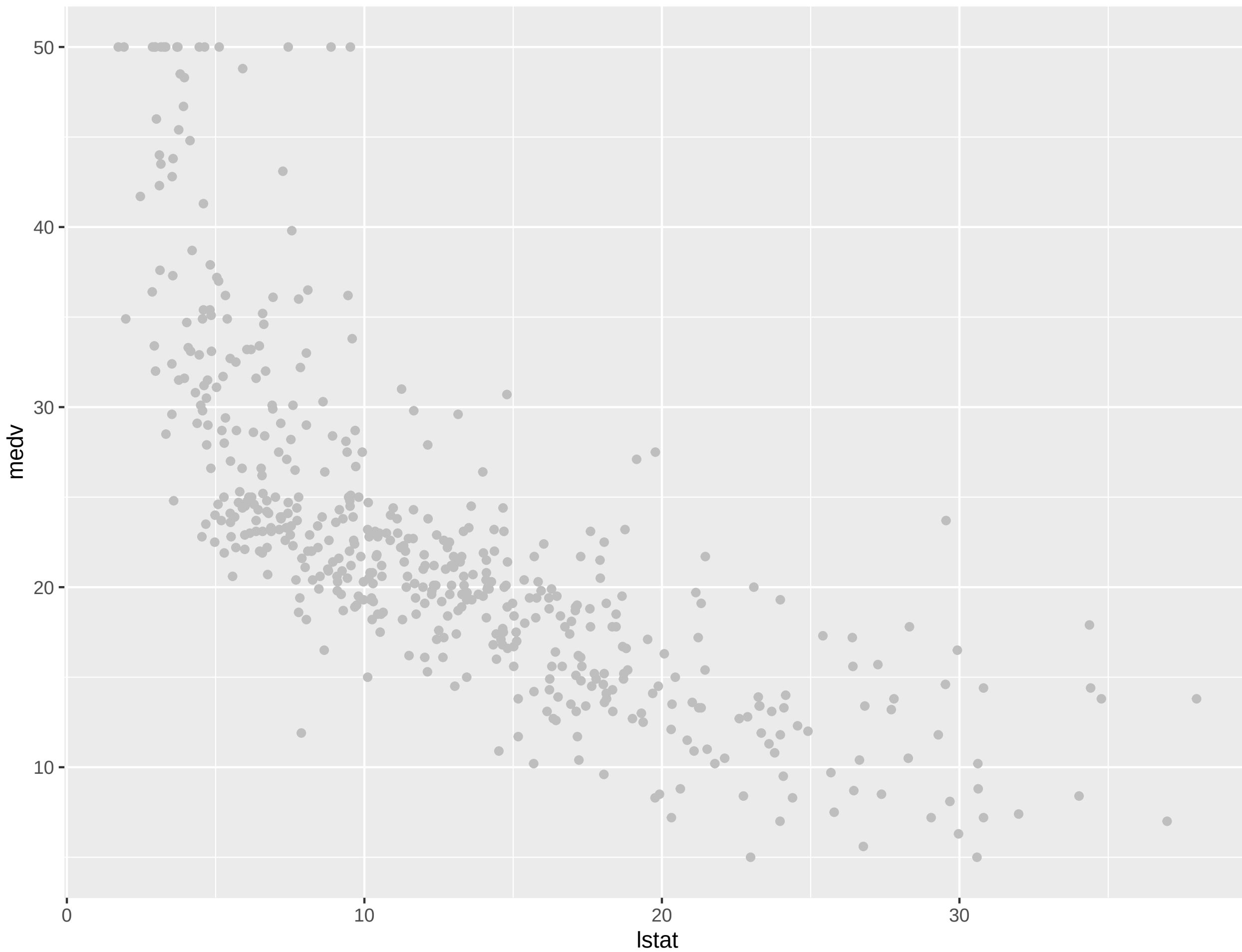
- Dependent variable: *medv*, the median value of owner-occupied homes (in thousands of dollars).
- Independent variable: *lstat* (percent of lower status of the population).

Load and plot the data

```
library(MASS)
data(Boston)
plot(Boston$lstat, Boston$medv, col = "gray", pch = 16,
     cex = 0.6, las = 1, xlab = "lower status of the population (percent)",
     ylab = "median value of owner-occupied homes")
```

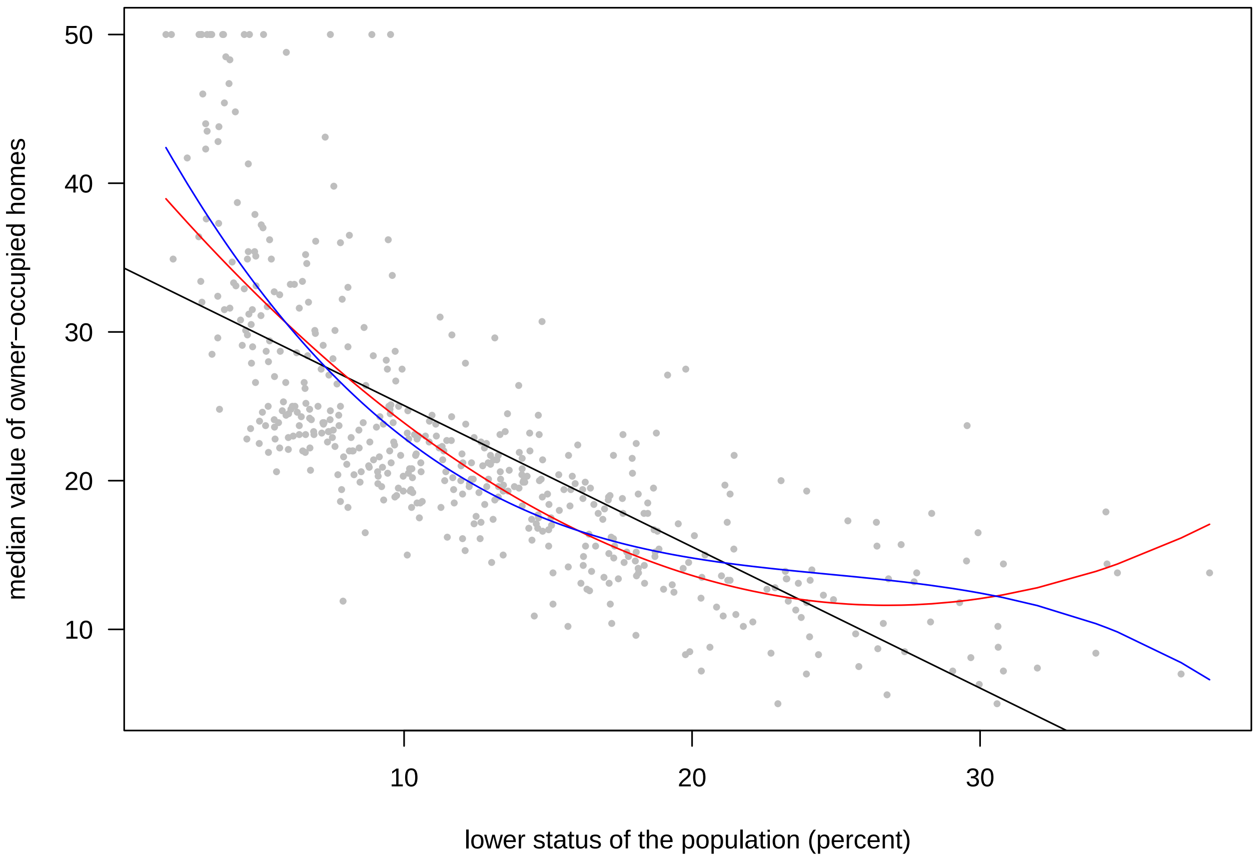



```
## ggplot
plot <- ggplot(aes(x = lstat, y = medv), data = Boston)
(plot <- plot + geom_point(colour = "gray"))
```



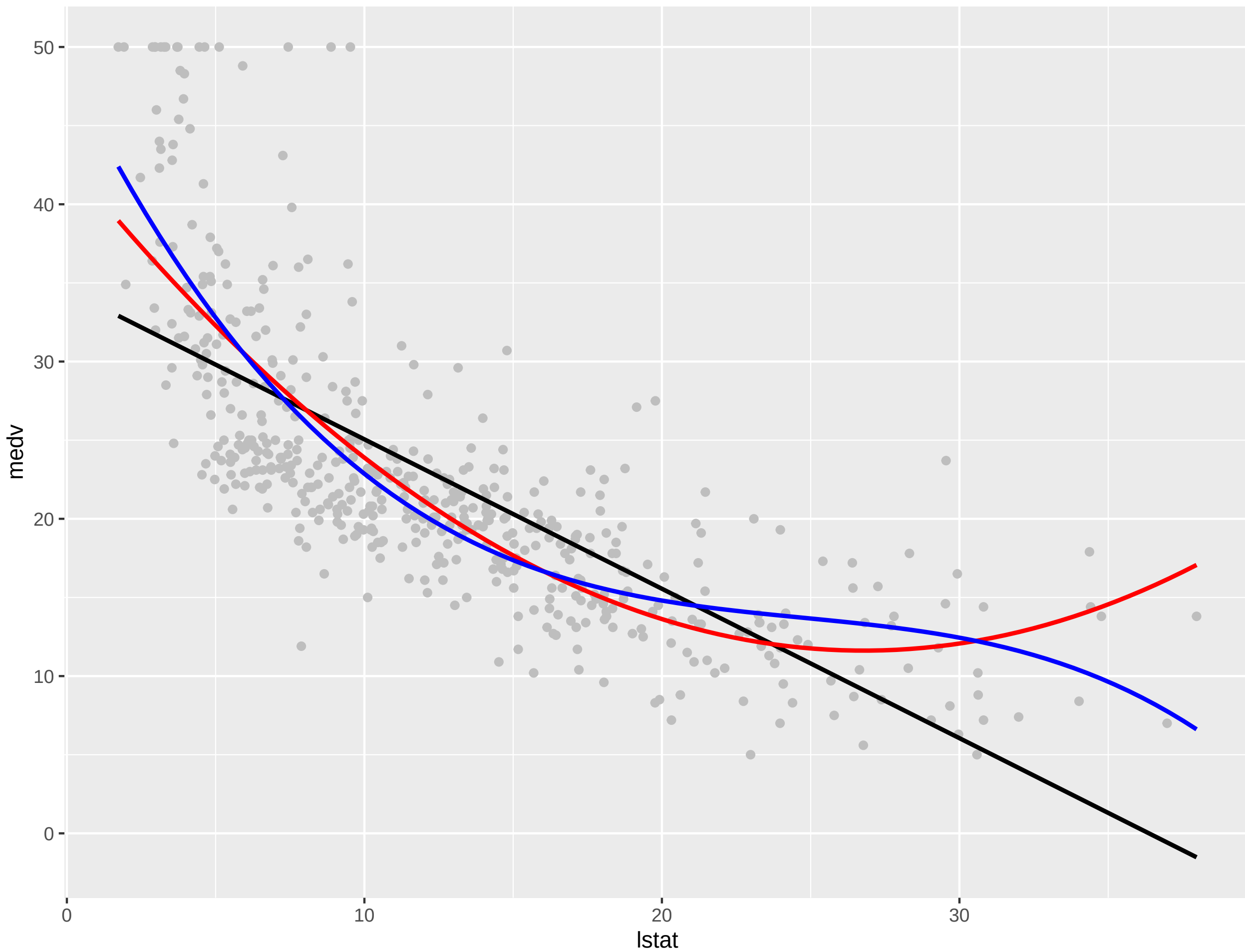
Plot the poylnomial regression fits

```
plot(Boston$lstat, Boston$medv, col = "gray", pch = 16,
     cex = 0.6, las = 1, xlab = "lower status of the population (percent)",
     ylab = "median value of owner-occupied homes")
## SLR
m1 <- lm(medv ~ lstat, data = Boston)
abline(m1)
## 2nd order polynomial fit
m2 <- lm(medv ~ lstat + I(lstat^2), data = Boston)
lines(sort(Boston$lstat), m2$fitted.values[order(Boston$lstat)], col = "red")
## 3rd order polynomial fit
m3 <- lm(medv ~ lstat + I(lstat^2) + I(lstat^3), data = Boston)
lines(sort(Boston$lstat), m3$fitted.values[order(Boston$lstat)], col = "blue")
```




```r
## Using ggplot
plot <- plot + geom_smooth(method = "lm", colour = "black", se = F)
plot <- plot + geom_smooth(method = "lm", formula = y ~ x + I(x^2), colour = "red", se = F)
plot <- plot + geom_smooth(method = "lm", formula = y ~ x + I(x^2) + I(x^3),
                           colour = "blue", se = F)
plot
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



ANOVA

```r
anova(M4)
```

```
## Analysis of Variance Table
##
## Response: Species
##            Df Sum Sq Mean Sq F value    Pr(>F)    
## Elevation   1 207828  207828 55.8981 1.023e-07 ***
## Area        1   3307    3307  0.8895 0.3550197    
## Adjacent    1  73171   73171 19.6804 0.0001742 ***
## Nearest     1   2909    2909  0.7823 0.3852165    
## Scruz       1   4636    4636  1.2469 0.2752082    
## Residuals 24  89231    3718                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Simulation

Step 1: Simulate the data sets

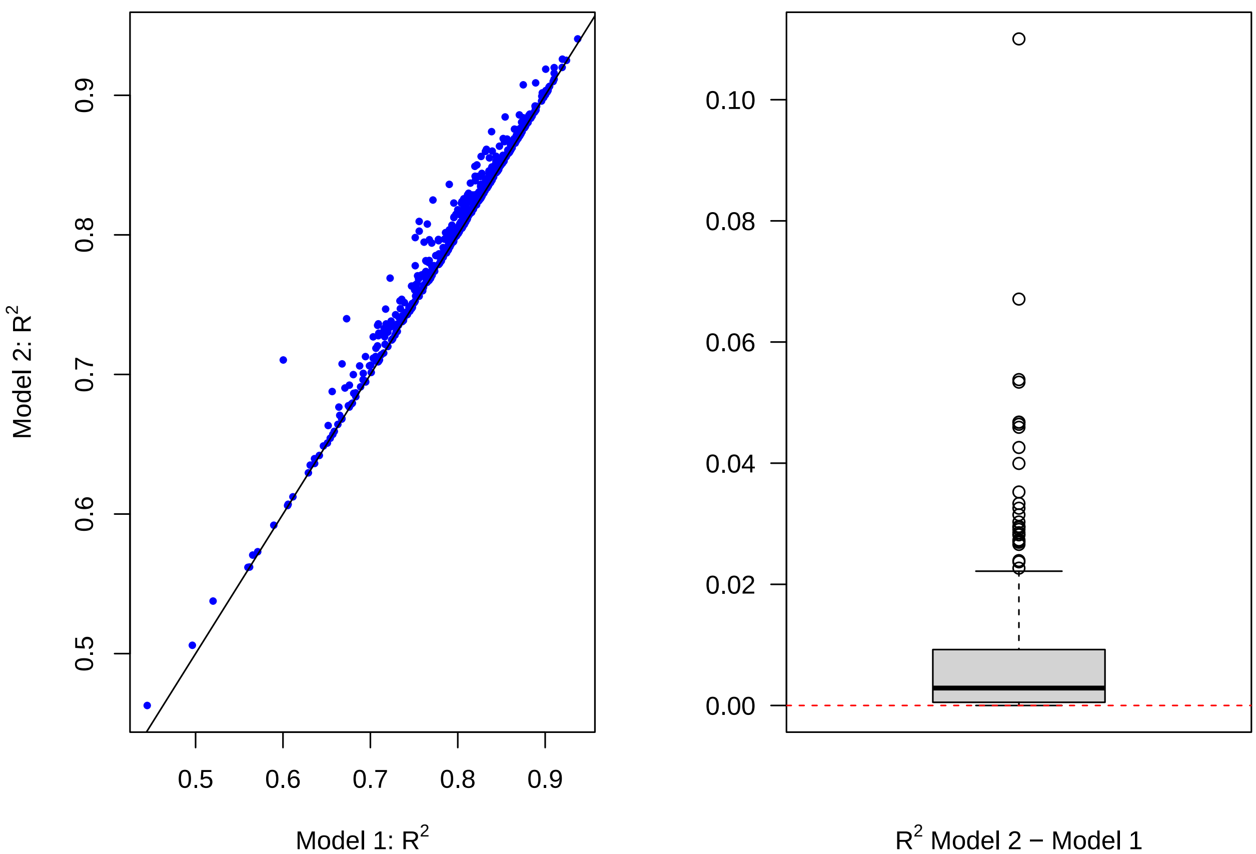
```
set.seed(123)
N = 500; n = 30
x1 <- replicate(N, rnorm(n))
x2 <- replicate(N, rnorm(n))
y1 <- apply(x1, 2, function(x) 5 + 2 * x + rnorm(n, 0, 1))
```

Step 2: Compute R^2 and R^2_{adj} for Model 1 and Model 2

```
R.sq <- array(dim = c(N, 4))
for (i in 1:N){
  R.sq[i, 1] = summary(lm(y1[, i] ~ x1[, i]))$r.squared
  R.sq[i, 2] = summary(lm(y1[, i] ~ x1[, i]))$adj.r.squared
  R.sq[i, 3] = summary(lm(y1[, i] ~ x1[, i] + x2[, i]))$r.squared
  R.sq[i, 4] = summary(lm(y1[, i] ~ x1[, i] + x2[, i]))$adj.r.squared
}
```

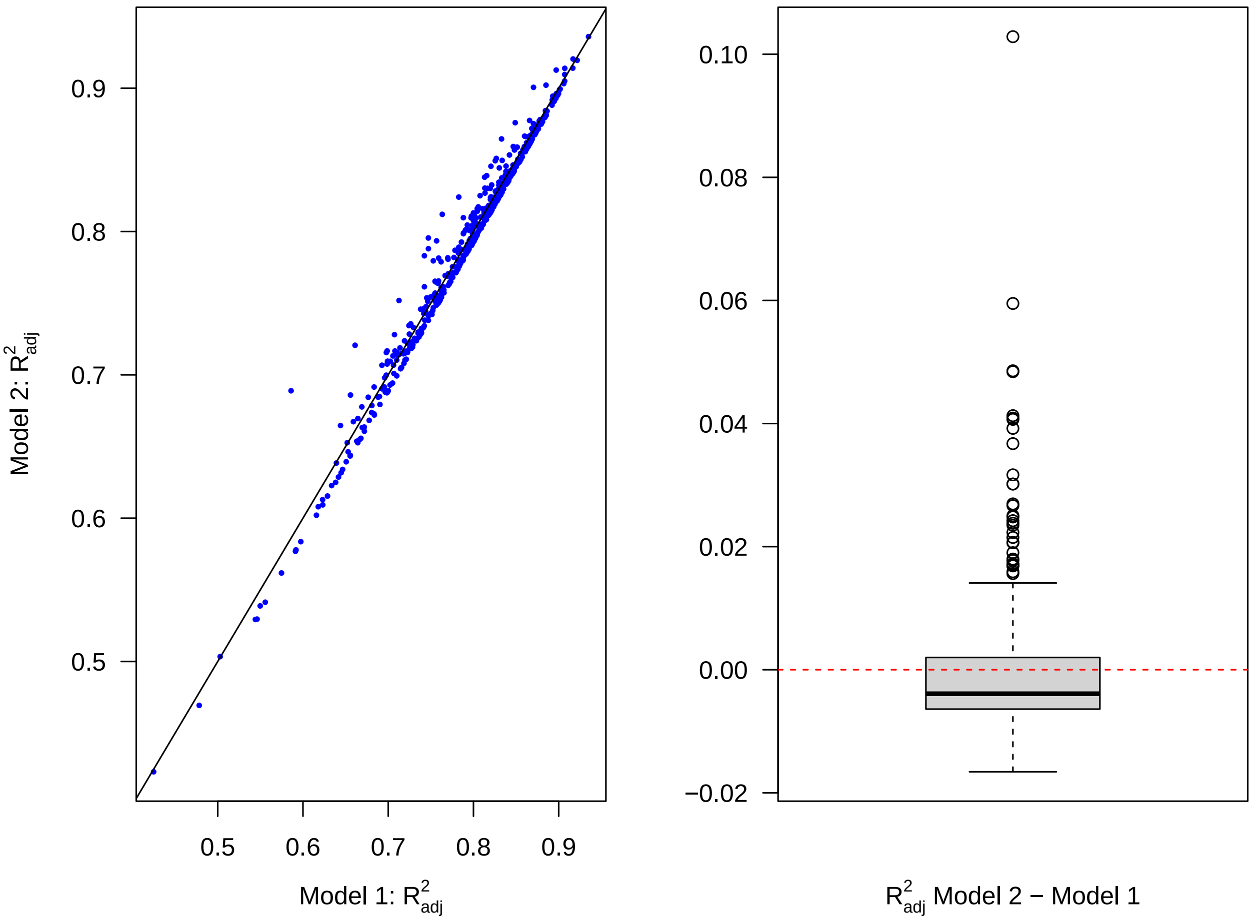
Compare R^2 for for Model 1 and Model 2

```
par(mfrow = c(1, 2))
plot(R.sq[, 1], R.sq[, 3], pch = 16, cex = 0.65, col = "blue",
     xlab = expression(paste("Model 1: ", R^2)),
     ylab = expression(paste("Model 2: ", R^2)))
abline(0, 1)

boxplot(R.sq[, 3] - R.sq[, 1], las = 1, xlab = expression(paste(R^2, " Model 2 - Model 1")))
abline(h = 0, lty = 2, col = "red")
```



Compare R^2_{adj} for for Model 1 and Model 2

```r
par(las = 1, mfrow = c(1, 2), mar = c(5.1, 4.6, 1.1, 1.1))
plot(R.sq[, 2], R.sq[, 4], pch = 16, cex = 0.5, col = "blue",
     xlab = expression(paste("Model 1: ", R[adj]^2)),
     ylab = expression(paste("Model 2: ", R[adj]^2)))
abline(0, 1)

boxplot(R.sq[, 4] - R.sq[, 2], las = 1, xlab = expression(paste(R[adj]^2, " Model 2 - Model 1")))
abline(h = 0, lty = 2, col = "red")
```



Model 2: R^2_{adj}
Model 1: R^2_{adj}
0.5
0.6
0.7
0.8
0.9
0.10
0.08
0.06
0.04
0.02
0.00
−0.02
R^2_{adj} Model 2 − Model 1