Quantitative covariates and regression analysis

Sonja Surjanovic
Applied Statistics and Data Science Group (ASDa)
Department of Statistics, UBC

January 29, 2018

Resources for statistical assistance

Department of Statistics at UBC:

www.stat.ubc.ca/how-can-you-get-help-your-data

SOS Program - An hour of free consulting to UBC graduate students. Funded by the Provost and VP Research Office.

STAT 551 - Stat grad students taking this course offer free statistical advice. Fall semester every academic year.

Short Term Consulting Service - Advice from Stat grad students. Fee-for-service on small projects (less than 15 hours).

Hourly Projects - ASDa professional staff. Fee-for-service consulting.

Outline

Correlation analysis

Simple linear regression

Multiple linear regression

ANOVA and ANCOVA

Methods for predicting continuous outcomes

The language of statistics is not as standardized as you might like!

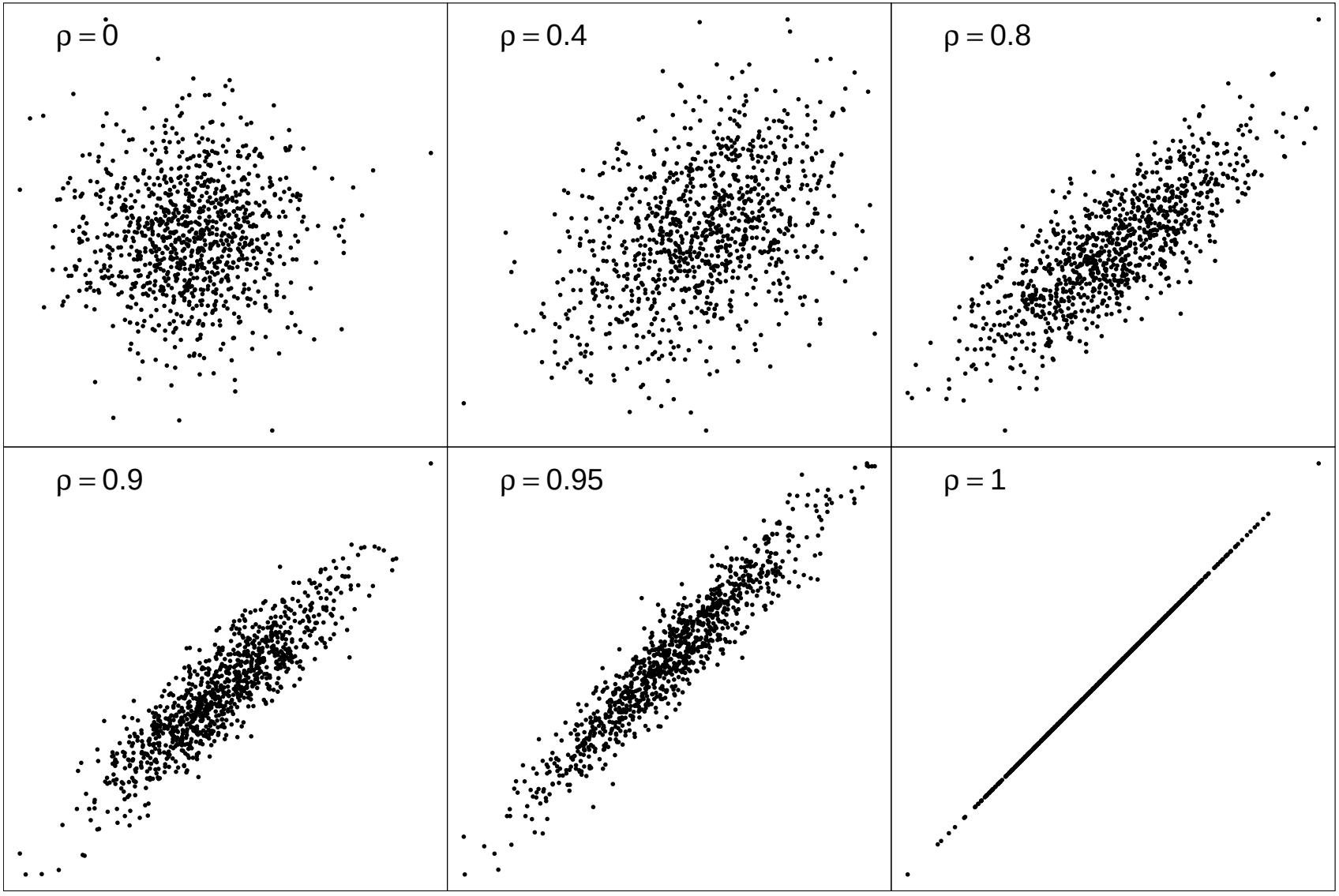
Different terms can be used for essentially the same model

Statisticians consider regression as the general approach

Method	**Type of predictor variable(s)**
Two sample t-test	1 categorical, 2 levels
ANOVA	1 or more categorical, 2 or more levels
Regression	1 or more continuous
ANCOVA	1 or more categorical and continuous

Correlation

Measures the direction and strength of relationship between numeric variables



Data format

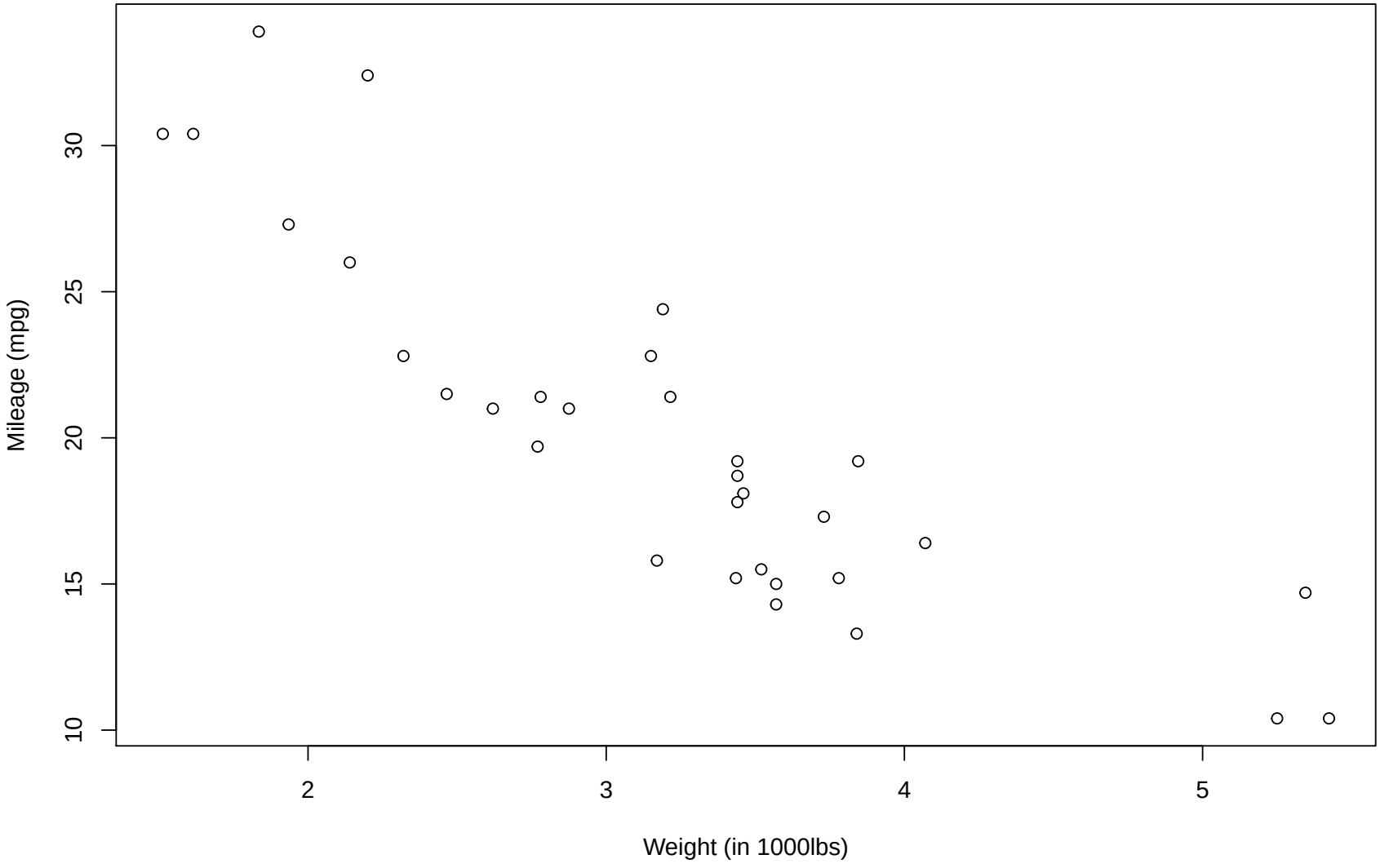
"Rectangular" data, fits in a rectangle where each row represents a sampled unit and each column is a characteristic observed on that unit

Example: Each row represents a different car model and the columns are various measured features

```
##                    mpg cyl    wt am gear
## Mazda RX4         21.0   6 2.620  1    4
## Mazda RX4 Wag     21.0   6 2.875  1    4
## Datsun 710        22.8   4 2.320  1    4
## Hornet 4 Drive    21.4   6 3.215  0    3
## Hornet Sportabout 18.7   8 3.440  0    3
## Valiant           18.1   6 3.460  0    3
## Duster 360        14.3   8 3.570  0    3
## Merc 240D         24.4   4 3.190  0    4
## Merc 230          22.8   4 3.150  0    4
## Merc 280          19.2   6 3.440  0    4
```

Association between car weight and mileage

1974 Motor Trends car data (32 different models of cars)



Summary statistics from the data

```
##        Mean Variance     SD Correlation
## mpg 20.091   36.324  6.027
## wt   3.217    0.957  0.978       -0.87
```

We can test the correlation between mpg and weight

```
##             Estimate t value  p value
## Correlation   -0.868   -9.56 1.29e-10

## 95% confidence interval:  -0.9338264 -0.7440872

## Squared Correlation =  0.753
```

Data (2 numeric variables)

From a random sample of n independent units from a population we measure 2 different variables (data)

Each variable has a mean and a variance and the two variables have a correlation

Let's call our variables X and Y

$\{X, Y\} = \{x_i, y_i\}$ for $i = 1, \ldots, n$

Correlation between x_i and x_j is 0

Correlation between y_i and y_j is 0

Correlation between x_i and y_i is ρ

Estimating the parameters from the sample

The population parameters:

- The means are denoted by μ_x and μ_y
- The variances are denoted by σ_x^2 and σ_y^2
- The covariance of the two variables is denoted by σ_{xy}
- The correlation is $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ is a number between -1 and 1

Formulas for the sample estimates of the population parameters:

$$\hat{\mu}_x = \bar{x} = \sum_1^n x_i / n$$

$$\hat{\sigma}_x^2 = s_x^2 = \sum_1^n (x_i - \bar{x})^2 / (n-1)$$

$$\hat{\sigma}_{xy} = s_{xy} = \sum_1^n (x_i - \bar{x})(y_i - \bar{y}) / (n-1)$$

$$\hat{\rho}_{xy} = r_{xy} = s_{xy} / (s_x s_y)$$

Correlation analysis

By correlation we usually mean the Pearson product-moment correlation coefficient. It measures the **linear** relationship between X and Y.

ρ is estimated by $r = \frac{s_{xy}}{s_x s_y}$

Hypothesis of interest is usually $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$ and we use the test statistic $r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}$

We can also use Fisher's Transformation ($F(r) = \frac{1}{2}\ln(\frac{1+r}{1-r})$) to test if ρ equals any value including 0 and to construct confidence intervals for ρ

Neither method is overly sensitive to the normality assumption. While $-1 \leq r \leq 1$, $-\infty < F(r) < \infty$

Spearman's rank correlation coefficient

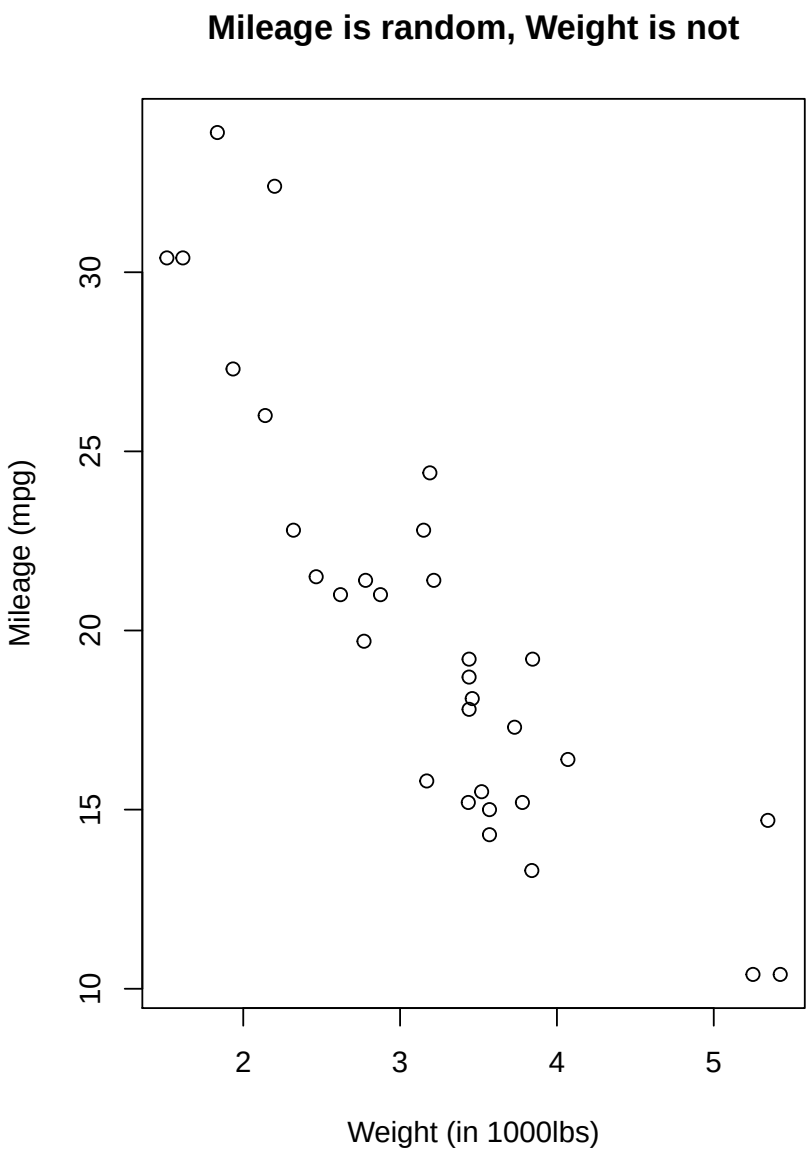
Is a rank version of Pearson's correlation computed by converting the data for each variable into a rank before computing the correlation. It uses same test statistics as Pearson's correlation.

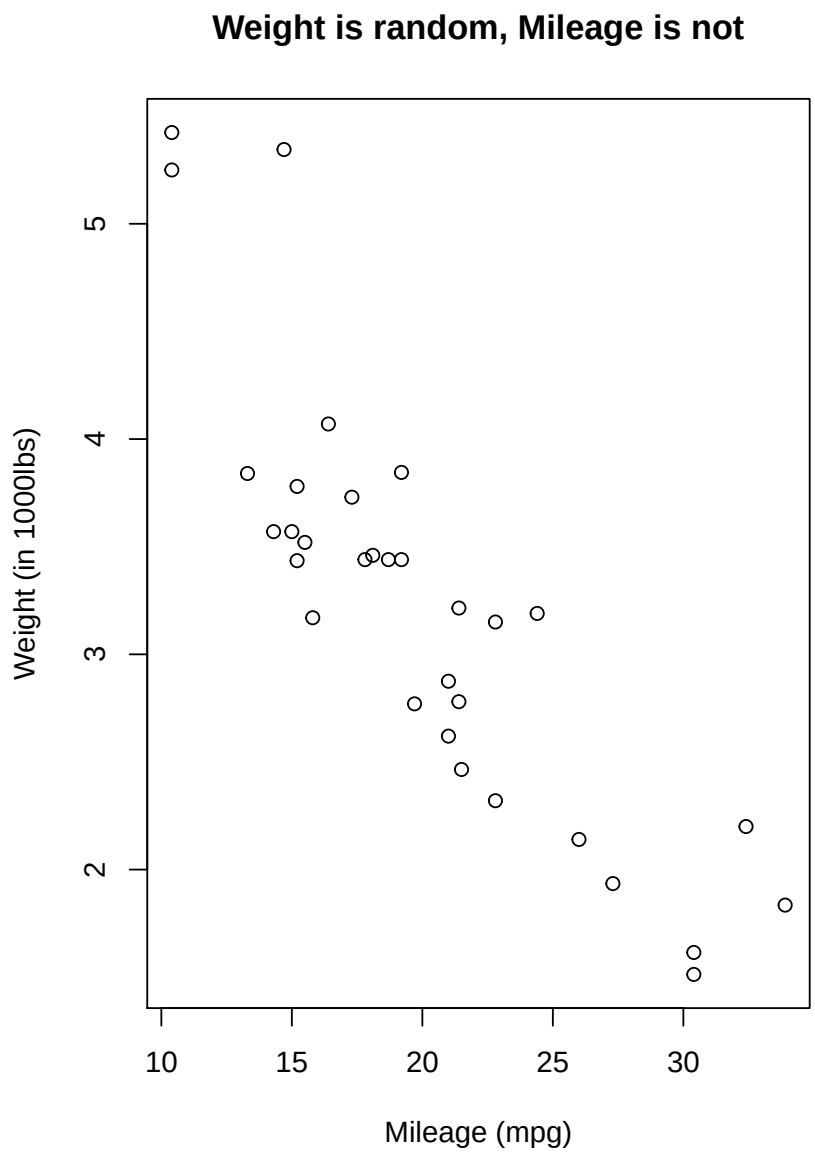
Is closely related to the Pearson's correlation coefficient except the relationship need not be linear in nature

Is robust to outliers

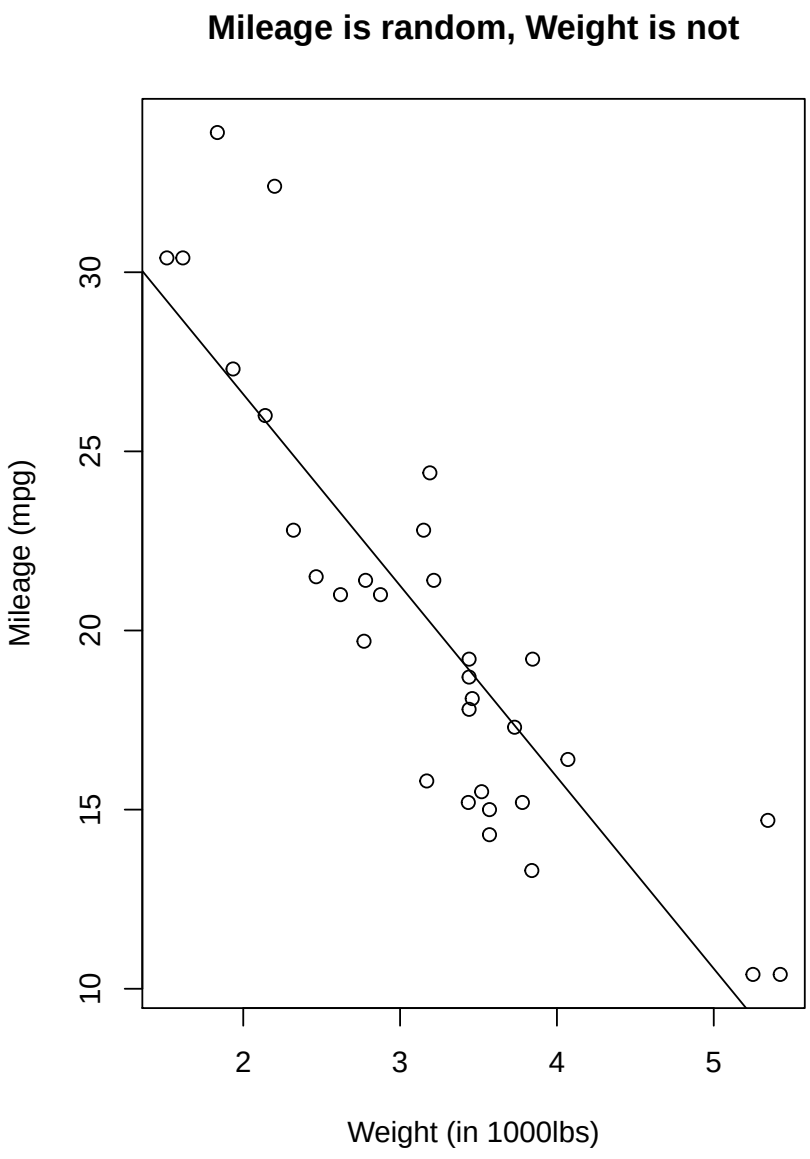
If the relationship is linear without any extreme points, Spearman's and Pearson's will be similar

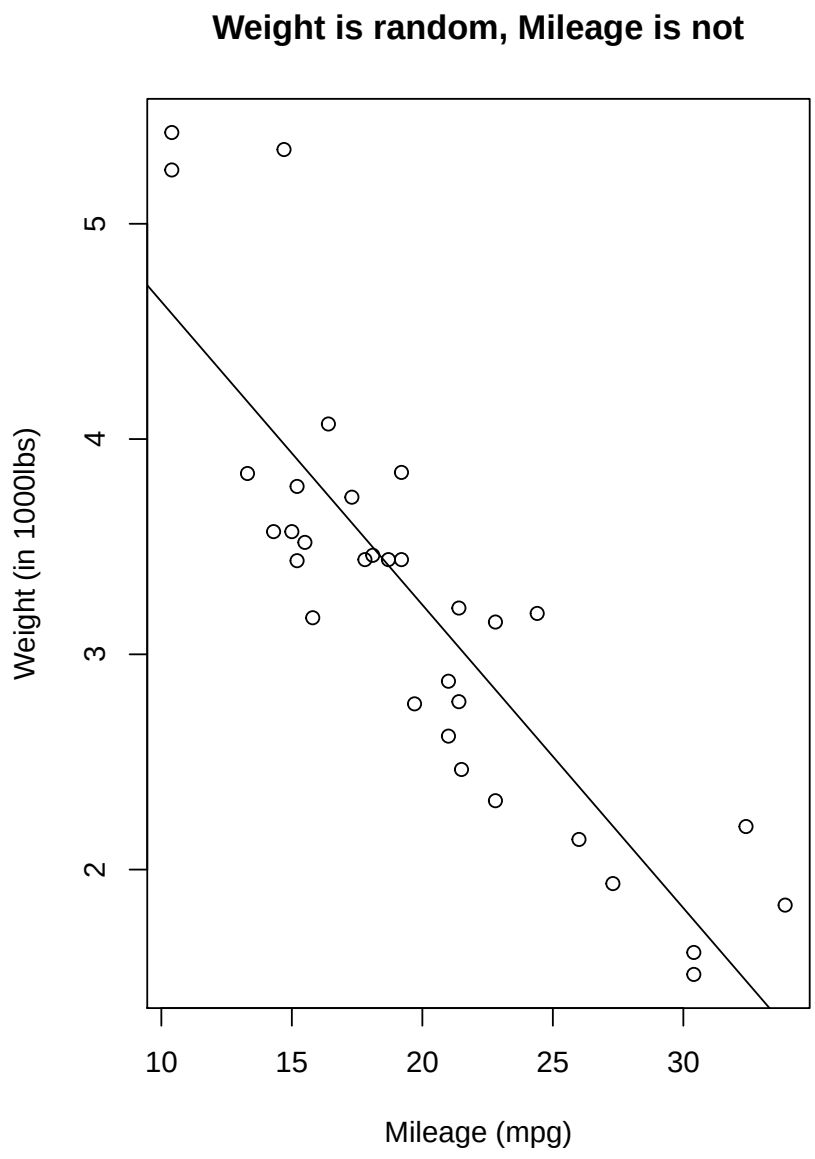
Two prediction scenarios





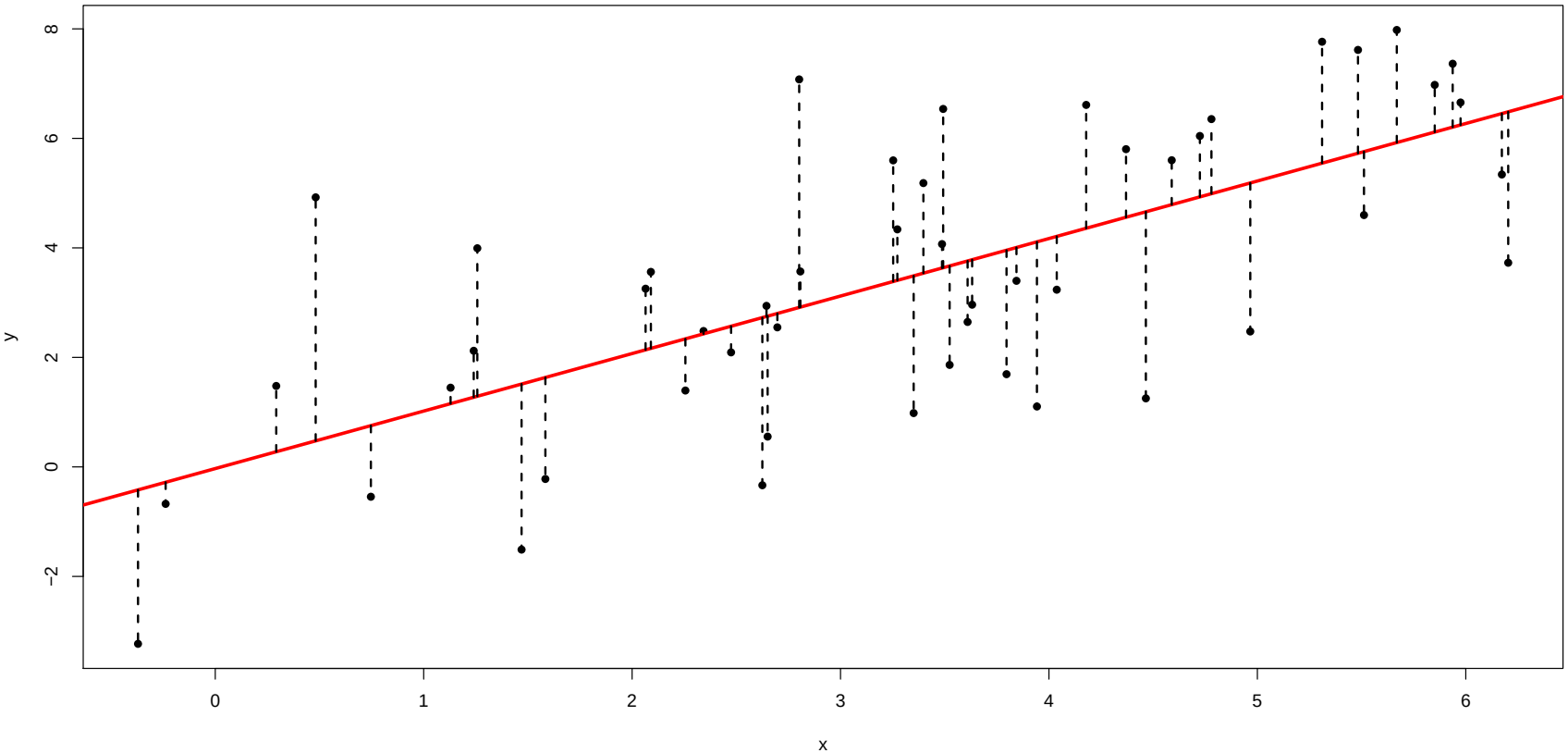
The prediction line





Determining the prediction line

Minimize the error (residuals), the vertical distance between the observed values and the predicted line



Minimize the sum of the squared errors, method called Least Squares

Two parameters determine the line: slope and intercept

Fitted values, residuals and mean squared error

We use the estimated slope and intercept with each x_i to compute a fitted value for y_i

$$\hat{y}_i = b_0 + b_1 x_i$$

We compute the residual by taking the difference between the observed data and the fitted value

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

We estimate the variance of the residuals

$$\hat{\sigma}^2_\varepsilon = s^2_\varepsilon = \sum \hat{\varepsilon}^2/(n-2)$$

Note: The average of the residuals is 0. We use $n-2$ because we estimated 2 parameters (b_0 and b_1).

Least Squares estimates for the slope and intercept

The slope is primarily determined by the correlation between Y and X. It's magnitude is limited by the ratio of the standard deviation of Y and X.

slope: $b_1 = r_{xy}\frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2}$

We estimate the intercept by picking a point on the line then using our estimate of the slope, solve for the intercept. $\{\bar{x}, \bar{y}\}$ is always on the line.

intercept: $b_0 = \bar{y} - b_1\bar{x}$

With some math it can be shown that both b_0 and b_1 can be computed as a weighted average of Y. This means both will follow a normal distribution if n is large enough by the central limit theorem of statistics (CLT).

Example revisited

Summary statistics from the data

```
##        Mean     SD Correlation
## mpg 20.091  6.027
## wt   3.217  0.978       -0.87
```

Regression coefficients predicting mpg by weight

```
## (Intercept)          wt
##       37.29       -5.34
```

```
## -0.868 * 6.027 / 0.978 = -5.344
## 20.091 - -5.344 * 3.217 = 37.285
```

Estimated line to predict mpg from weight

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    37.29      1.878   19.86 8.24e-19
## wt             -5.34      0.559   -9.56 1.29e-10

## R squared = 0.753
```

Estimated line to predict weight from mgp

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.047     0.3087   19.59 1.20e-18
## mpg           -0.141     0.0147   -9.56 1.29e-10

## R squared = 0.753
```

Compare the 3 results

```
##             Estimate t value  p value
## Correlation   -0.868   -9.56 1.29e-10
```

```
##    Estimate Std. Error t value Pr(>|t|)
## wt    -5.34      0.559   -9.56 1.29e-10
```

```
##     Estimate Std. Error t value Pr(>|t|)
## mpg   -0.141     0.0147   -9.56 1.29e-10
```

All three have the same t value and p value

Fitted values and residuals

$\hat{y}_i = b_0 + b_1 x_i$

$Fit_i = 37.29 + (-5.34 wt_i)$

$\hat{\varepsilon}_i = y_i - \hat{y}_i$

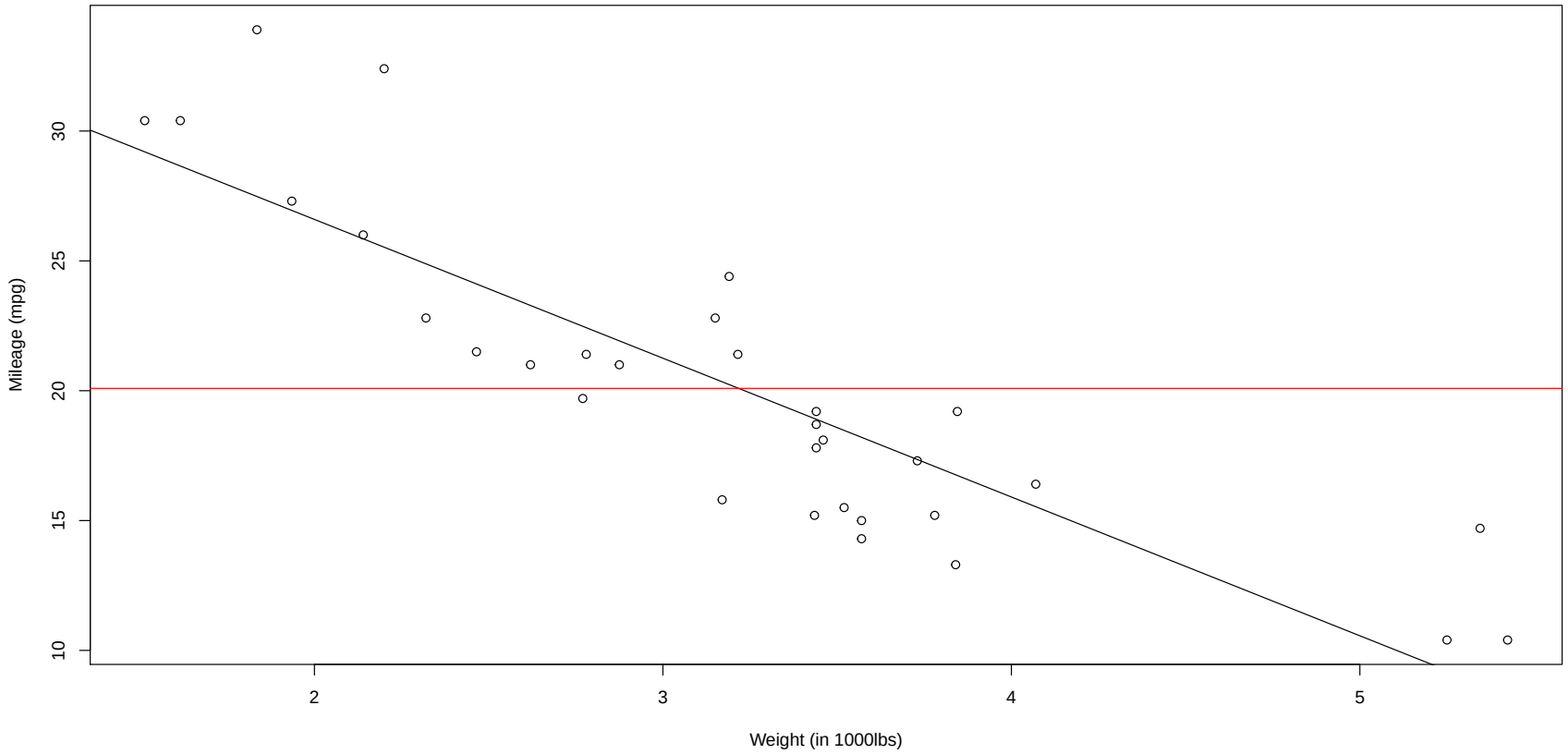
$Res_i = wt_i - Fit_i$

```
##                     wt  mpg  Fit   Res
## Mazda RX4         2.62 21.0 23.3 -2.28
## Mazda RX4 Wag     2.88 21.0 21.9 -0.92
## Datsun 710        2.32 22.8 24.9 -2.09
## Hornet 4 Drive    3.21 21.4 20.1  1.30
## Hornet Sportabout 3.44 18.7 18.9 -0.20
```

```
## MSE =  9.277
```

The coefficient of determination

Better prediction with regression line compared to red mean line?



Variation line doesn't explain (sum square of errors to line) = 278

Total variation in y (sum square of errors to $\bar{y}$) = 1126

% variation NOT explained by the line = 278 / 1126 = 0.247

% variation explained by the line = 1 - 0.247 = 0.753

The coefficient of determination

Tells how much better at predicting Y the regression model is compared to just using $\bar{y}$

- proportion of the variance of Y explained by the model
- depends on the size of the errors (residuals)

For simple linear regression $R^2 = \rho_{xy}^2$

The R^2 alone doesn't tell the whole story:

- The more variables in the model the higher the R^2 BUT also the higher the variability in the predictions made from the model (due to having to estimate more coefficients for the model)
- With large sample sizes, the R^2 value could still be low even with highly significant model coefficients
- With small sample sizes, the R^2 value could still be high even with insignificant model coefficients

The error of the fitted value

Decreases as x_i moves closer to $\bar{x}$

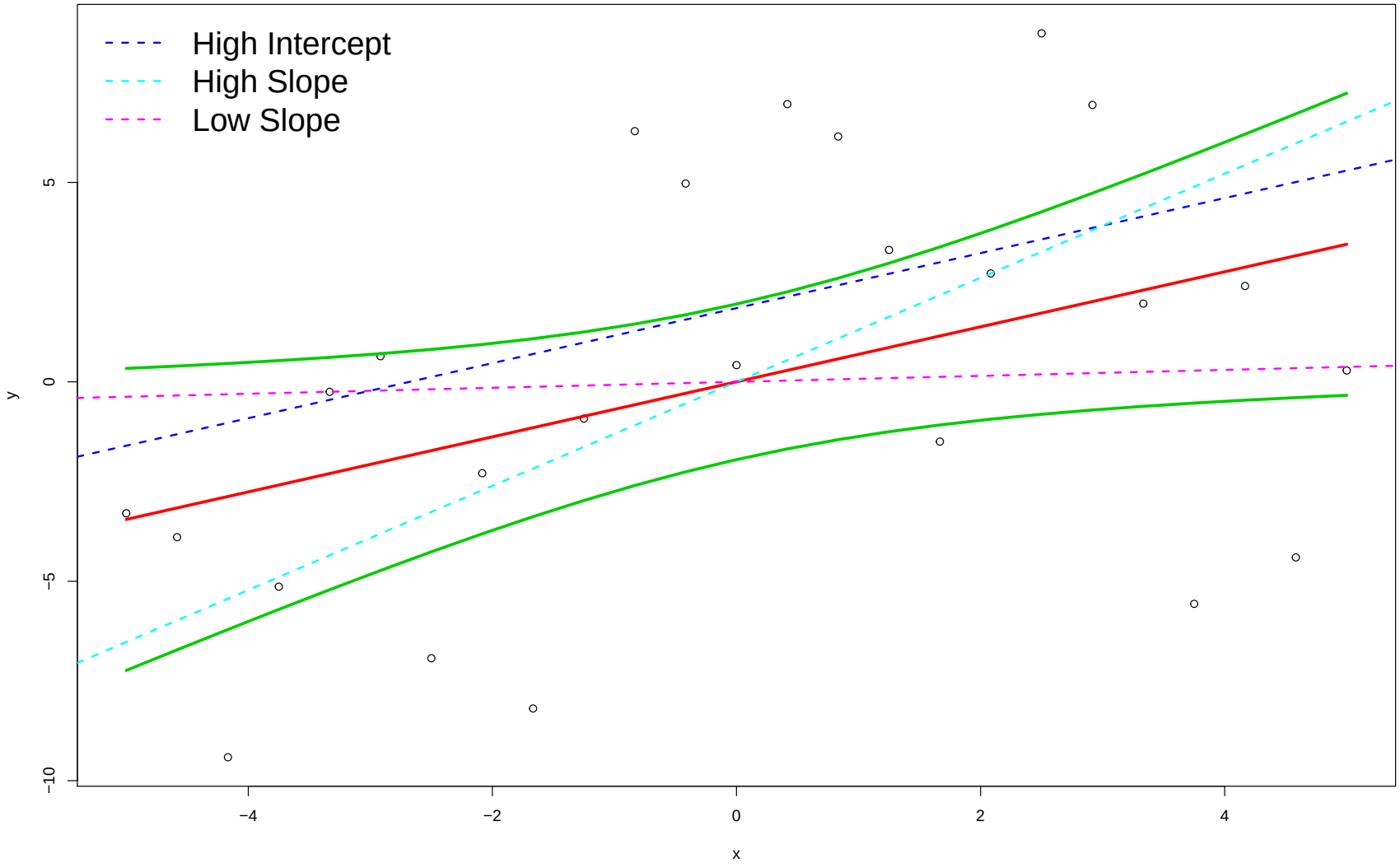
- fitted values at the endpoints depend greatly on the slope of the line
- fitted values at the middle are relatively insensitive to the slope

Decreases as the variance in x increases

- since the slope is determined by the endpoints

The distribution of the fitted value can be approximated by a normal distribution because of the CLT. This means we can compute accurate confidence bounds for the fitted line.

95% Confidence Interval (green) for the fitted line



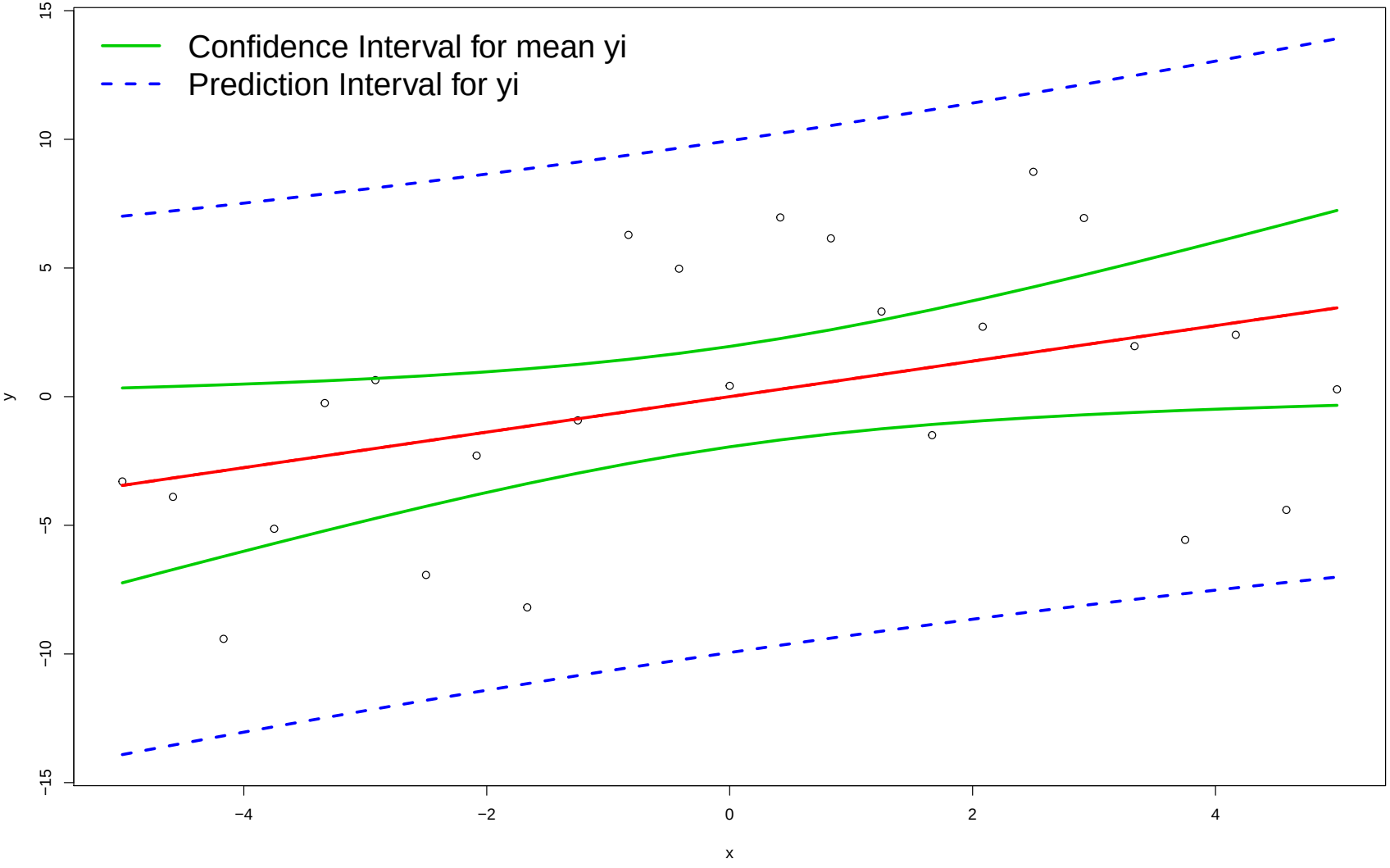
There is a 95% chance that the interval at a given x_i covers the true mean for y_i

Prediction Interval

The prediction estimate is the same as the fitted value

The uncertainty (variability) in the prediction includes the uncertainty in the fitted line (model) plus the variability in the y data

95% Prediction Interval (blue dashed)



There is a 95% chance that the interval at a given x_i covers the true y_i

Simple Linear Regression

A line that summarizes the relationship between two quantitative variables and is used to make predictions

Regression assumes that Y is random but X is not

Model the mean of Y as a function of X

$\mu_{y_i} = \beta_0 + \beta_1 x_i$

Each observation is described as being some distance (error ε_i) from the estimated mean where $\varepsilon_i \sim N(0, \sigma^2)$

Assumptions:

- normality is not critical
- homoscedasticity (constant variance) is important
- independent errors is critical

Diagnostics for Regression

Before fitting the model plot y_i versus x_i

- Does the spread of Y depend on X?
- What does the relationship look like?
- Are there any extreme X or Y values?

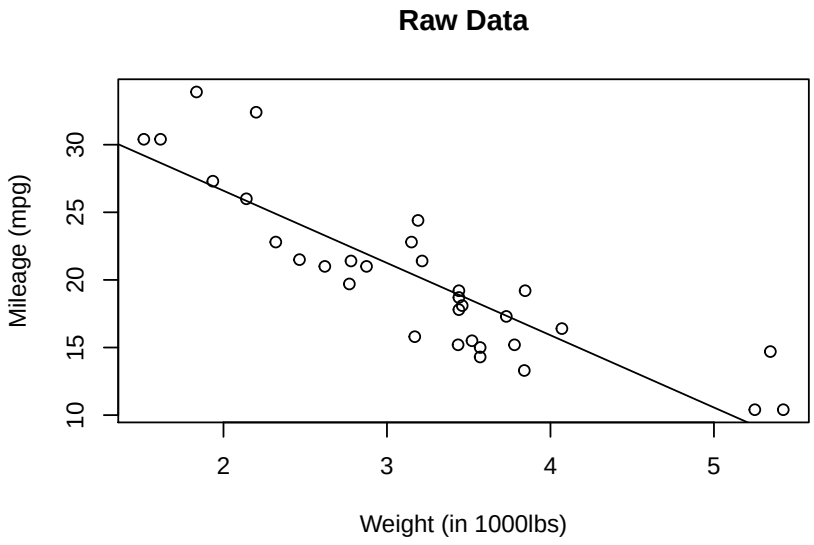
Plot the residuals versus the quantiles of a normal distribution

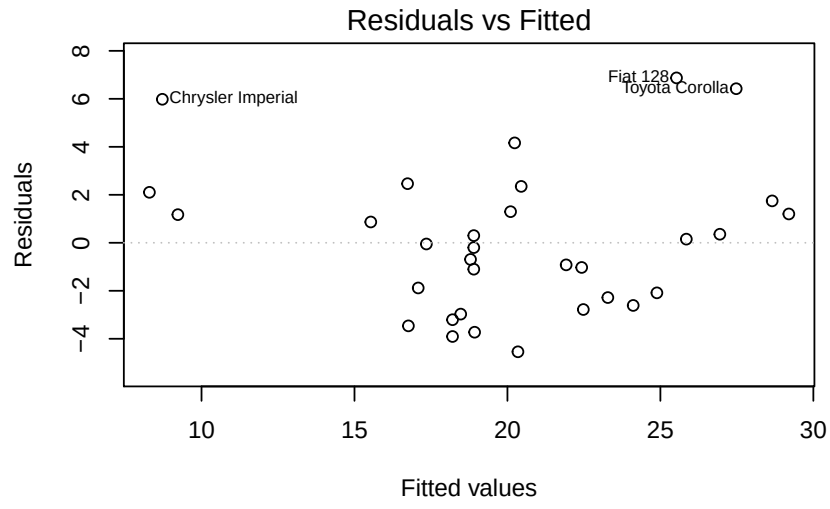
Plot residuals (ε_i) versus fitted values ($\hat{y}_i$). You should see an uncorrelated oval of data.

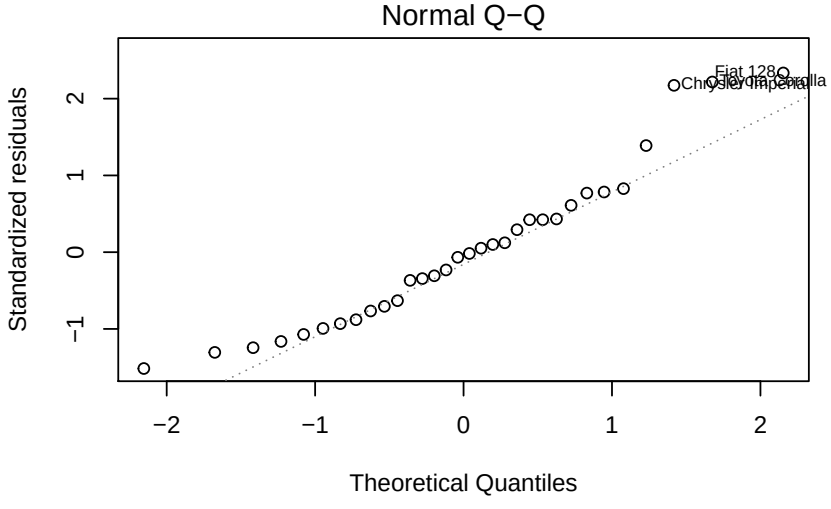
If the data can be ordered over time or space, check the residuals for indications of serial correlation

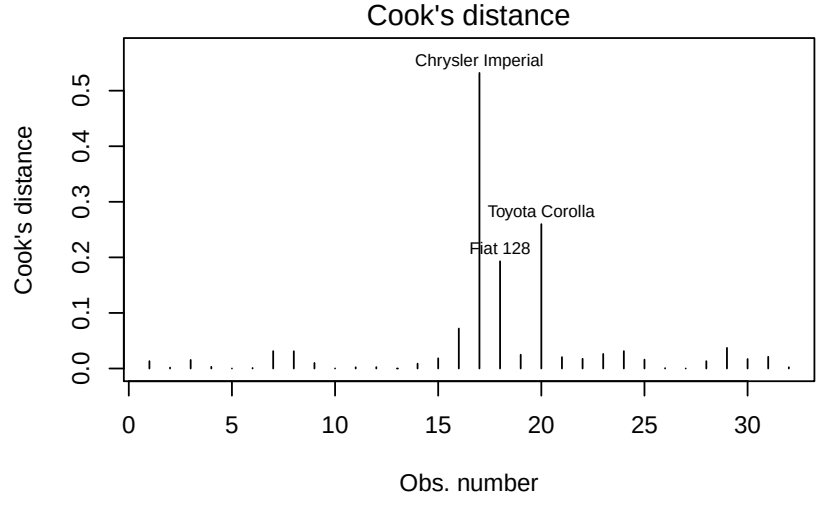
Examine the influence of each observation using Cook's distance

Example Revisited

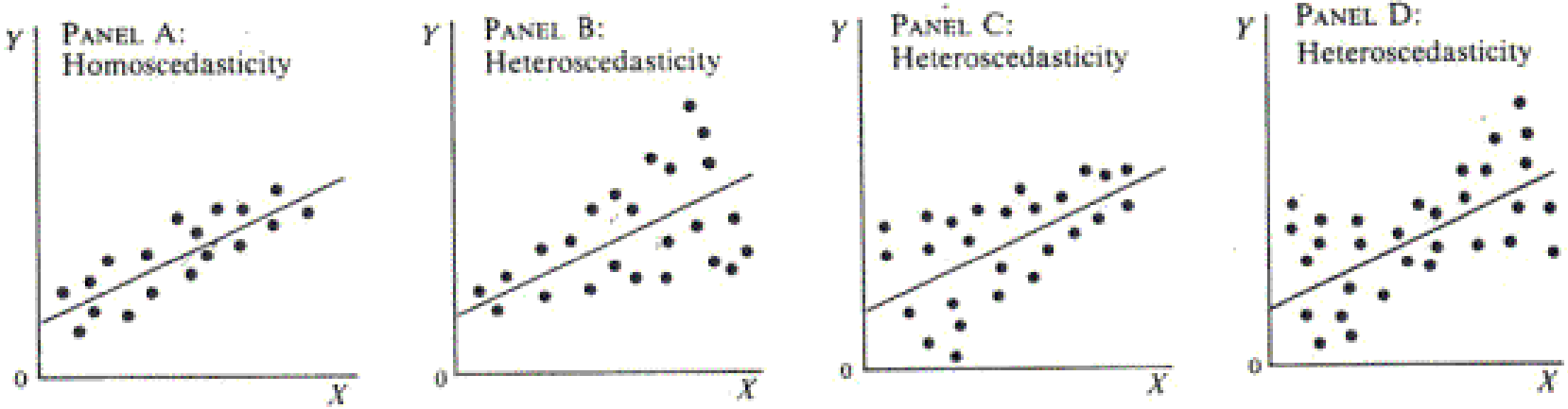




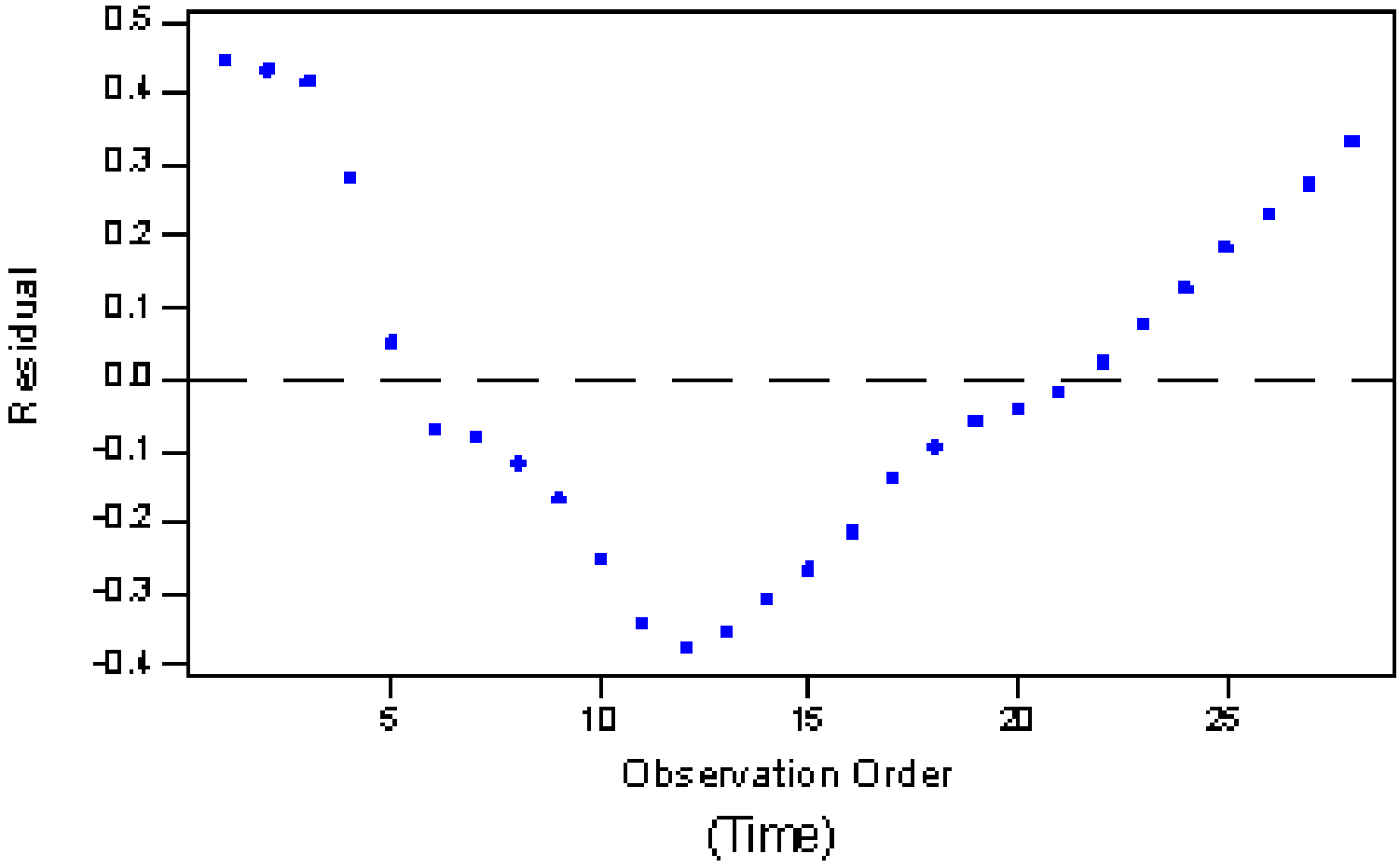




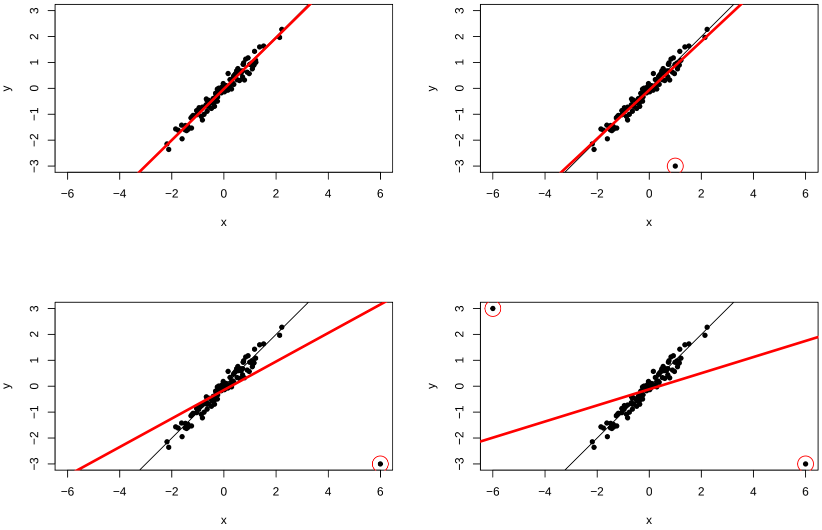
Heteroscedasticity



Serial Correlation

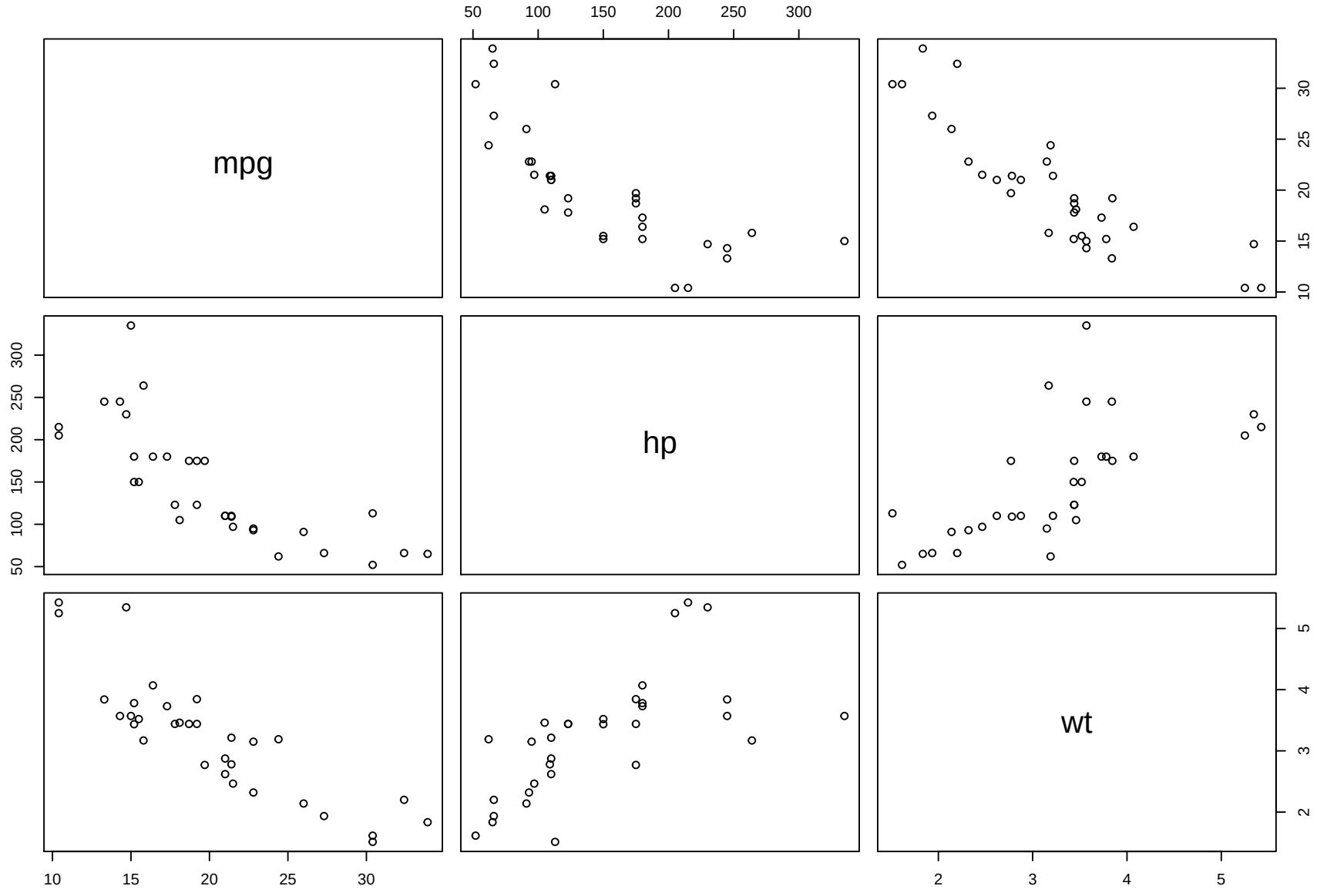


Effect of Outliers or high leverage points



Regession with 2 predictors

1974 Motor Trends car data (32 different models of cars)



Summary statistics from the data

```
##          Mean Variance    SD
## mpg    20.091   36.324  6.03
## hp    146.688 4700.867 68.56
## wt      3.217    0.957  0.98
```

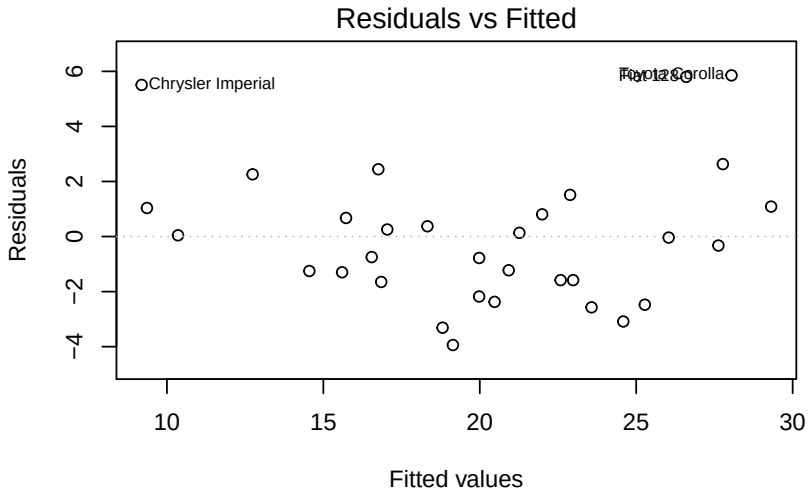
```
## correlations between the variables
```

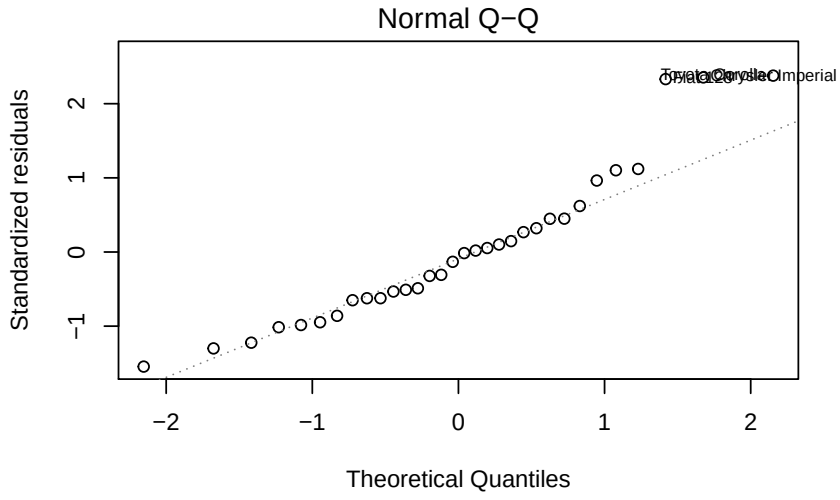
```
##         mpg     hp     wt
## mpg   1.000 -0.776 -0.868
## hp   -0.776  1.000  0.659
## wt   -0.868  0.659  1.000
```

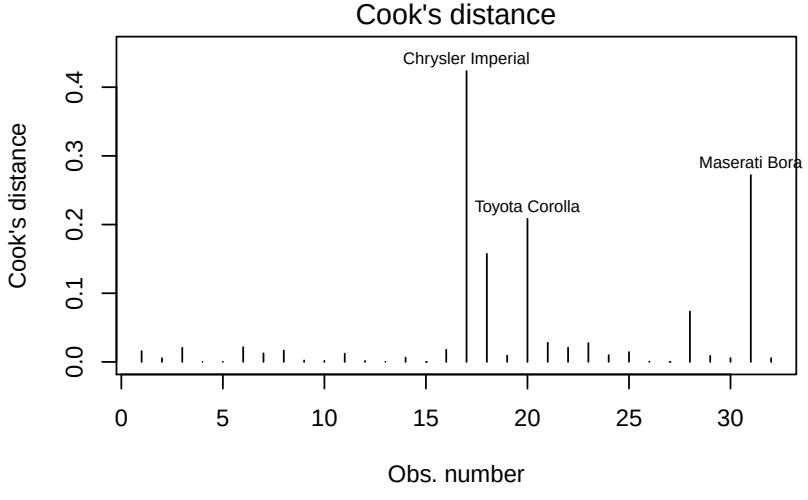
Regression predicting mpg by hp and weight (wt)

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2273    1.59879   23.28 2.57e-20
## wt           -3.8778    0.63273   -6.13 1.12e-06
## hp           -0.0318    0.00903   -3.52 1.45e-03

## R squared = 0.827

##                   wt  hp  mpg  Fit   Res
## Mazda RX4       2.62 110 21.0 23.6 -2.57
## Mazda RX4 Wag   2.88 110 21.0 22.6 -1.58
## Datsun 710      2.32  93 22.8 25.3 -2.48

## MSE =  6.726
```

Model Diagnostics







Multiple Linear Regression (2 predictors)

The assumptions are the same but the equation (model) contains more parameters:

- $\mu_Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- computing the estimates of the coefficients is more complicated but can be easily expressed in matrix form (not presented here)

The key concern is how much correlation exists between X_1 and X_2 since severe multicollinearity can increase the variance of the coefficient estimates

- can complicate or prevent the identification of an optimal set of explanatory variables for a statistical model
- assess using the variance inflation factor (VIF), $1/(1 - \rho^2_{x_1,x_2})$
- if VIF $>= 5$, coefficients are poorly estimated and one should be wary of their p-values
- doesn't affect how well the model fits, a model with severe multicollinearity can produce great predictions

Example 1 Random data Y, X_1, X_2

True model: $Y = X_1$ with $\rho_{x_1,x_2} = 0$

```
##    Estimate Std. Error
## X1 1.005275 0.03165875
```

```
##         Estimate Std. Error
## X2 -0.007381078 0.03193062
```

```
##    Estimate Std. Error
## X1  1.00527    0.03166
## X2 -0.00010    0.03177
```

SE multiplier = 1.0

Example 2 Random data Y, X_1, X_2

True model: $Y = X_1$ with $\rho_{x_1,x_2} = 0.5$

```
##    Estimate Std. Error
## X1 1.005082 0.03172337
```

```
##     Estimate Std. Error
## X2 0.4973381 0.03194161
```

```
##         Estimate Std. Error
## X1  1.005911683 0.03649992
## X2 -0.001682867 0.03661223
```

SE multiplier $= 0.0365/0.0317 = 1.15$

Example 3 Random data Y, X_1, X_2

True model: $Y = X_1$ with $\rho_{x_1,x_2} = 0.9$

```
##    Estimate Std. Error
## X1 1.004437 0.03179358
```

```
##     Estimate Std. Error
## X2 0.9027353 0.03187386
```

```
##         Estimate Std. Error
## X1  1.010369666 0.07246802
## X2 -0.006612264 0.07258049
```

SE multiplier $= 0.0725/0.0318 = 2.28$

Example 4 Random Data Y, X_1, X_2

True model: $Y = X_1/2 + X_2/2$ with $\rho_{x_1,x_2} = 0$

```
##      Estimate Std. Error
## X1 0.5016772 0.03169791
```

```
##      Estimate Std. Error
## X2 0.4962426 0.03181129
```

```
##      Estimate Std. Error
## X1 0.5052747 0.03165973
## X2 0.4999045 0.03177184
```

Example 5 Random data Y, X_1, X_2

True model: $Y = X_1/2 + X_2/2$ with $\rho_{x_1,x_2} = 0.5$

```
##      Estimate Std. Error
## X1 0.7516066 0.03175274
```

```
##    Estimate Std. Error
## X2 0.749294 0.03185154
```

```
##      Estimate Std. Error
## X1 0.5059117 0.03649992
## X2 0.4983171 0.03661223
```

Example 6 Random Data Y, X_1, X_2

True model: $Y = X_1/2 + X_2/2$ with $\rho_{x_1,x_2} = 0.9$

```
##      Estimate Std. Error
## X1 0.9530507 0.03180093
```

```
##      Estimate Std. Error
## X2 0.9527279 0.03185082
```

```
##      Estimate Std. Error
## X1 0.5103697 0.07246802
## X2 0.4933877 0.07258049
```

Dealing with multicollinearity

Possible solutions:

- remove highly correlated predictors
- linearly combine predictors (add them together)
- do nothing if the p-values aren't important
- standardize the predictors (subtract the mean)

If you can live with less precise coefficient estimates, or a model that has a high R-squared but few significant predictors, doing nothing can be the correct decision because it won't impact the fit

What makes a regession linear?

Linear regression refers to the coefficients in the model. It has nothing to do with the way X is included in the model or the relationship between X and Y. It is the form of β:

Linear regression models:

$$\mu_y = \beta_0 + \beta_1 X + \beta_2 X^2$$

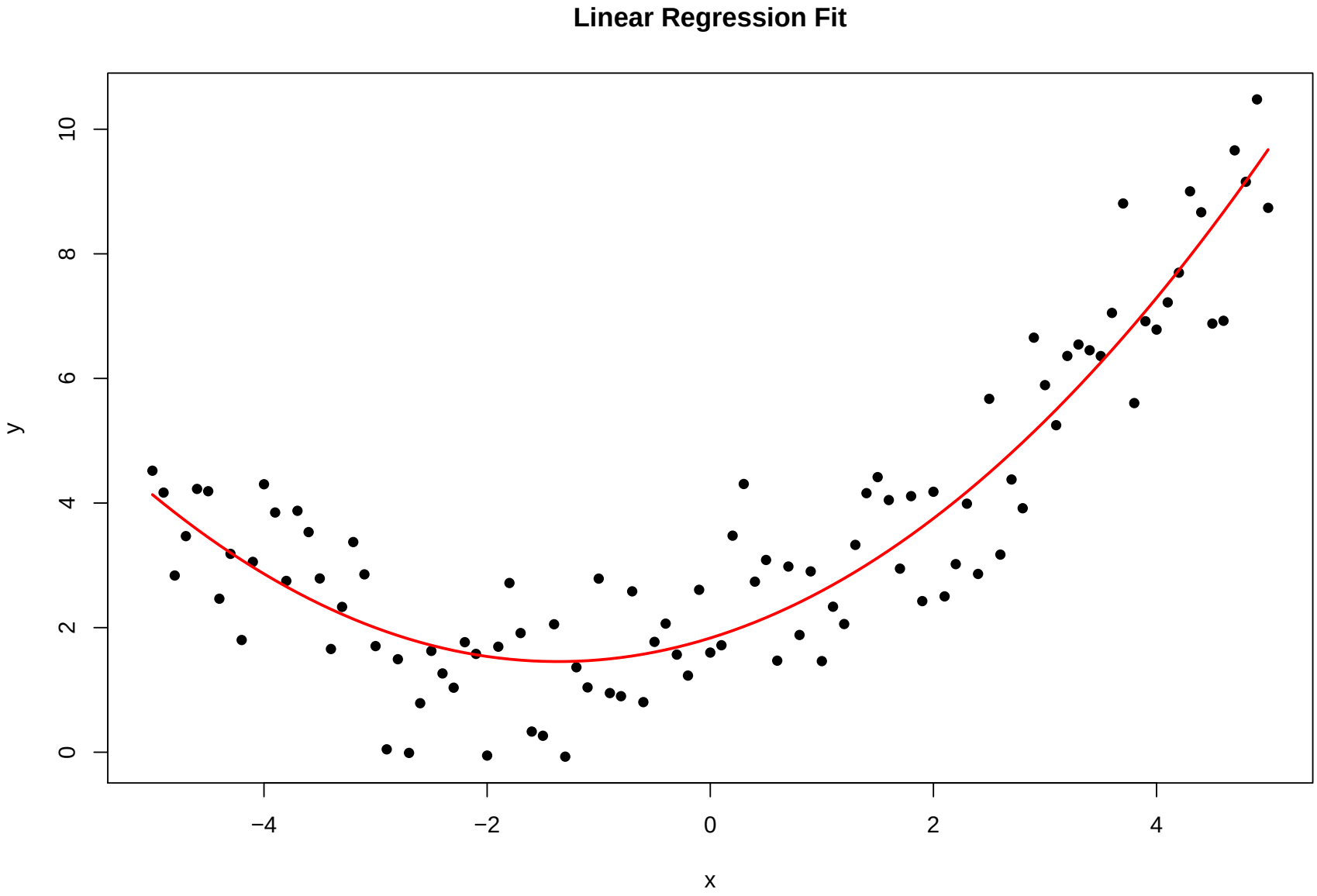
$$\mu_y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_1^2 + \beta_5 X_2^2$$

Not a linear regression model because of β_1^2:

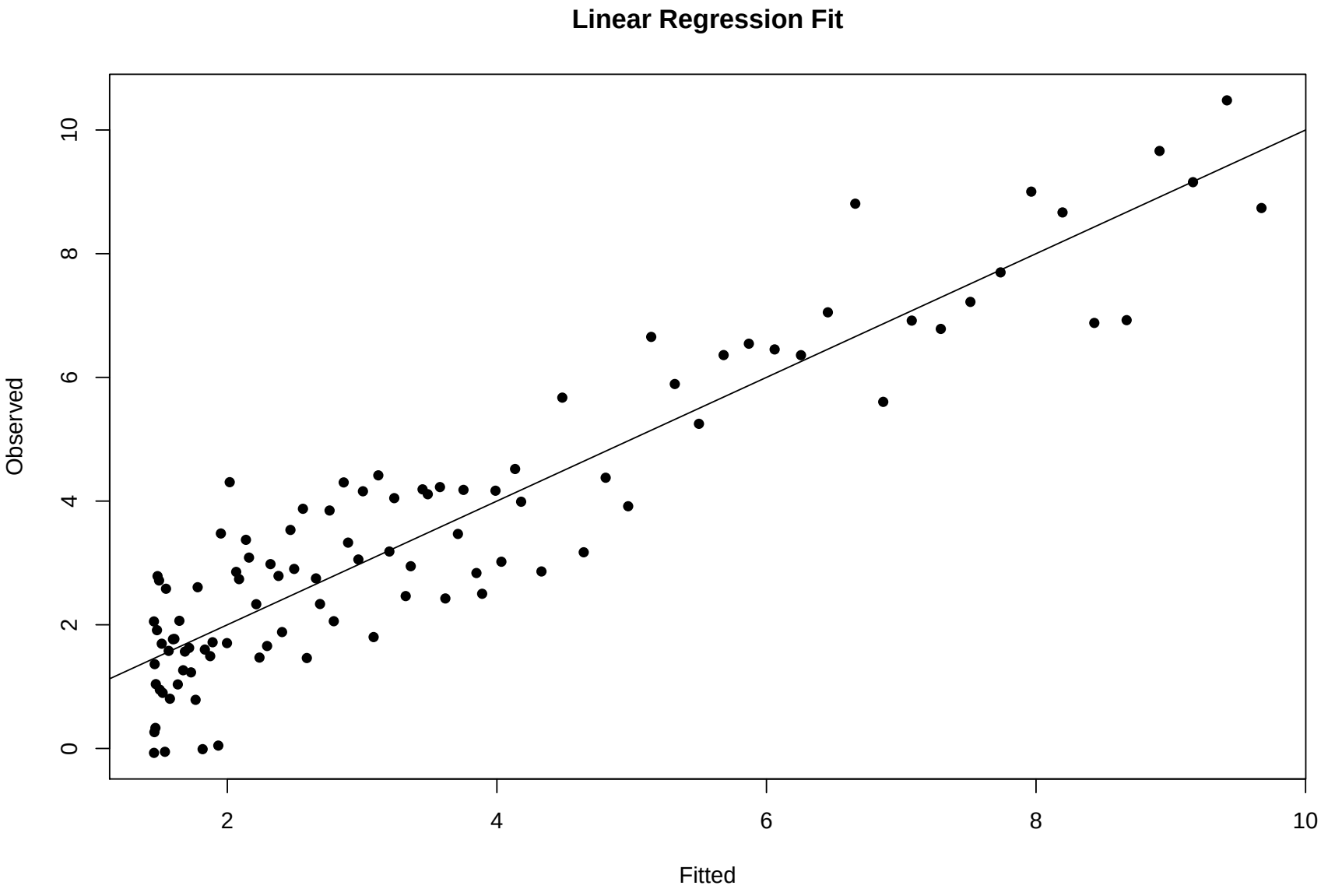
$$\mu_y = \beta_0 + \beta_1 X_1 + \beta_1^2 X_2$$

This is a (linear) Regression

$$\mu_y = 2 + x/2 + x^2/5$$



The observed versus the model fit



Regression with k predictors

Like the two predictor case, we can compute all the parameter estimates easily using matrix algebra. The model includes more terms but otherwise the issues are similar to the two predictor case.

We need to look at the correlation matrix of all the predictors to determine if colinearity is a problem. While a predictor may not be highly correlated to any one other predictor, it may be highly correlated with a set of other predictors.

If more than two predictors are included in the model then the VIF is unique for each predictor and is computed by regressing each predictor on the other predictors in the model and looking at the coefficient of determination

ANOVA as a regression model (categorical predictors)

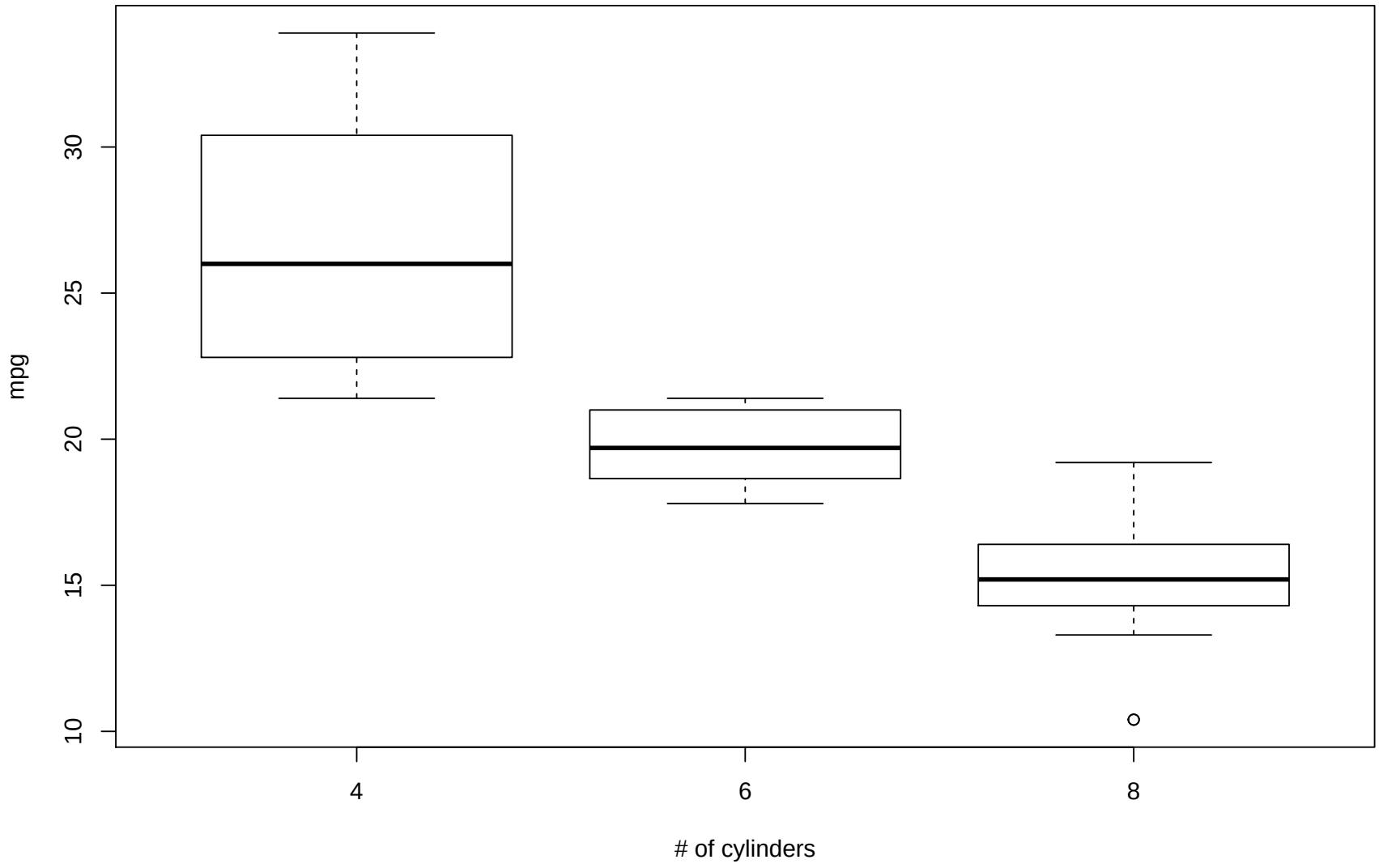
In order to fit an ANOVA model as a regression model, the p level factor variable is converted into $p - 1$ indicator variables

Example: The 3 level cylinder predictor is converted into 2 indicator variables, one for 6 cylinders and one for 8 cylinders

```
##                    mpg cyl
## Mazda RX4         21.0   6
## Mazda RX4 Wag     21.0   6
## Datsun 710        22.8   4
## Hornet 4 Drive    21.4   6
## Hornet Sportabout 18.7   8
```

```
##                    mpg cyl6 cyl8
## Mazda RX4         21.0    1    0
## Mazda RX4 Wag     21.0    1    0
## Datsun 710        22.8    0    0
## Hornet 4 Drive    21.4    1    0
## Hornet Sportabout 18.7    0    1
```

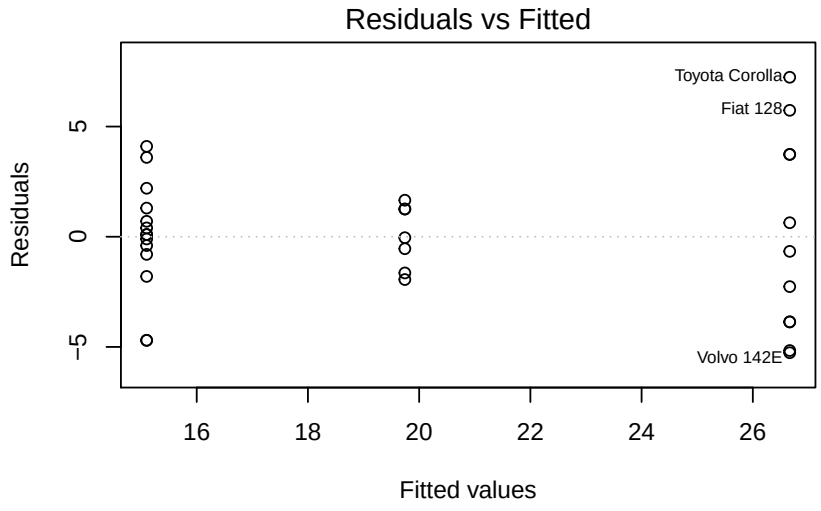
Example

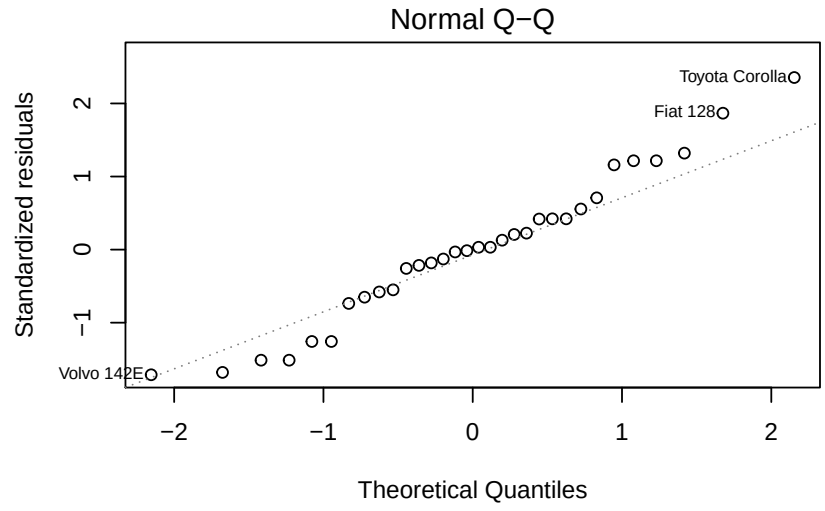


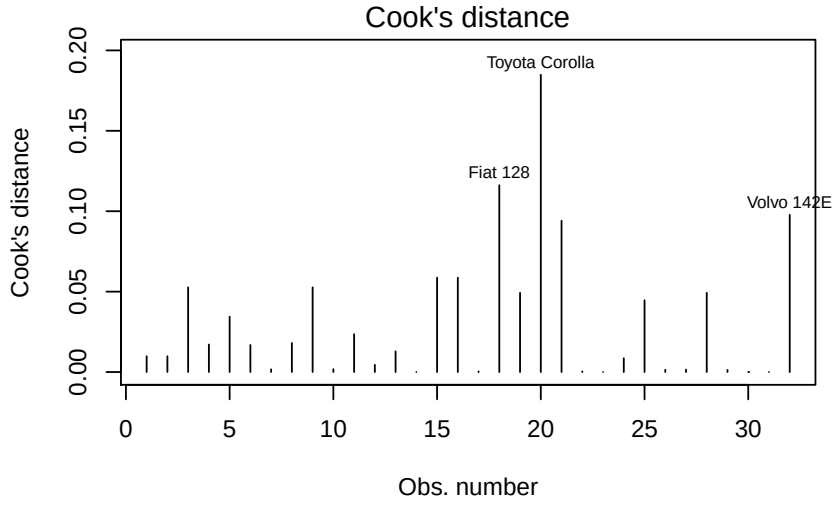
```
##   cyl  N  Mean    SD
## 1   4 11 26.66 4.510
## 2   6  7 19.74 1.454
## 3   8 14 15.10 2.560
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)       26.66      0.972   27.44 2.69e-22
## cyl6              -6.92      1.558   -4.44 1.19e-04
## cyl8             -11.56      1.299   -8.90 8.57e-10
```

What about the constant variance assumption?







ANCOVA continuous and categorical predictor

This is the common names for a model that contains both categorical and continuous predictors. We include categorical predictors by converting them into indicator variables then proceed with a multiple regression. Only difference is this time the model also includes a continuous predictor X

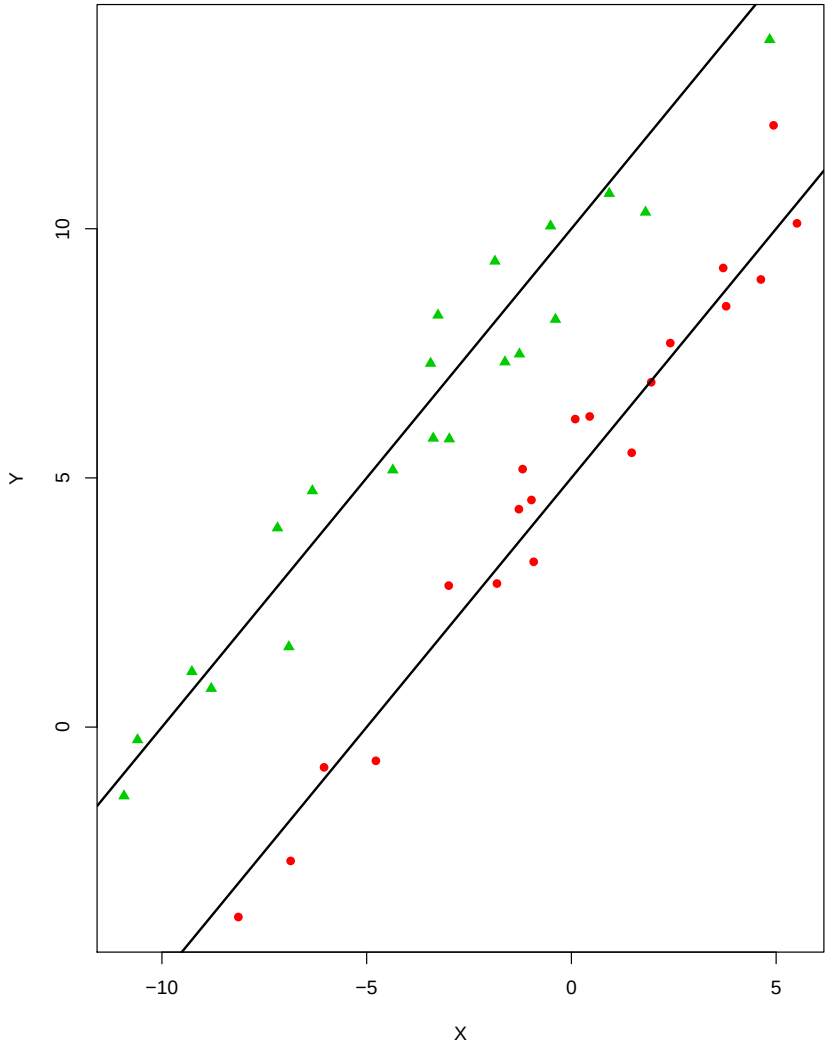
Example: Treatment versus placebo with a covariate

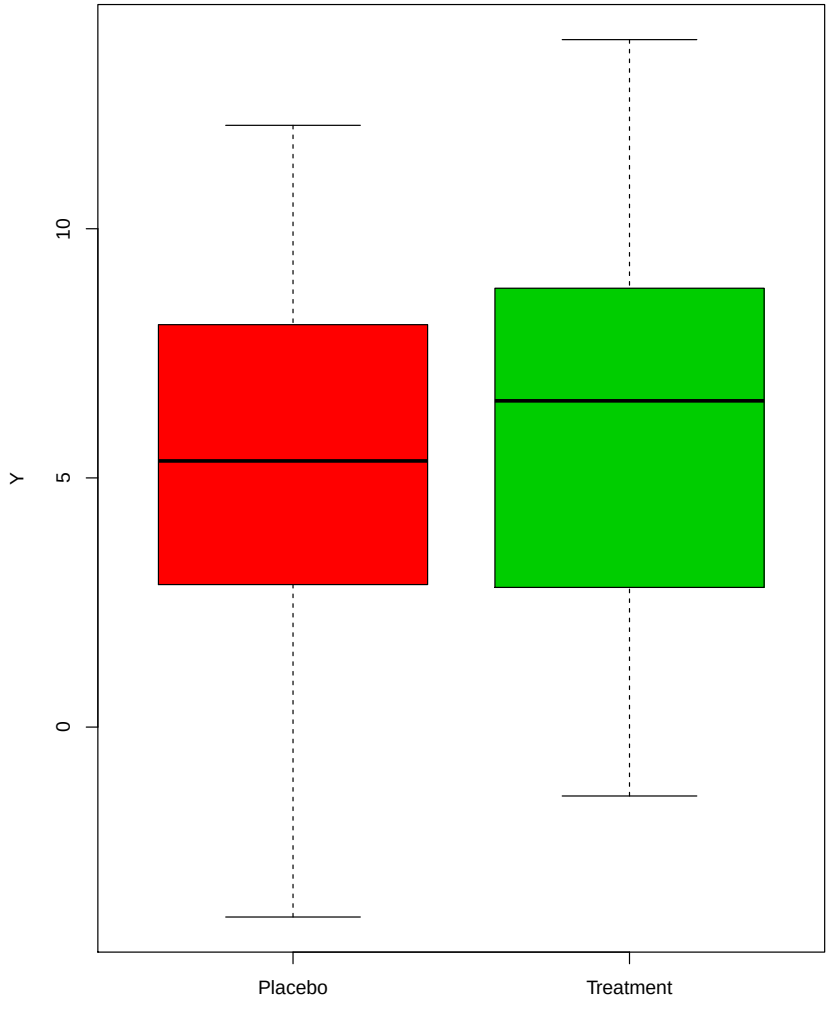
$$\mu_Y = \beta_0 + \beta_1 I(trt) + \beta_2 X$$

$b_0 =$ intercept for the placebo group
$b_1 =$ change in intercept from placebo to treatment
$b_2 =$ common slope for the covariate

This model fits parallel lines

$$\mu_y = 5 + 5I(trt) + X$$





Regression estimates

```
##          Estimate Std. Error
## Placebo 4.826374  0.9359718
## TRT-PLB 1.179350  1.3236640
```

```
##             Estimate Std. Error
## Intercept 7.0034277 0.41183350
## Slope     0.7787446 0.08507816
```

```
##           Estimate Std. Error
## Placebo 5.1247779  0.2127760
## TRT-PLB 4.6019328  0.3271369
## Slope   0.9859269  0.0372547
```

Model with different slopes

The slopes of our continuous predictor need not be the same for each level of our categorical predictor. We can model this with an interaction term.

$$\mu_Y = \beta_0 + \beta_1 I(trt) + \beta_2 X + \beta_3 X I(trt)$$

b_0 = intercept for the placebo group

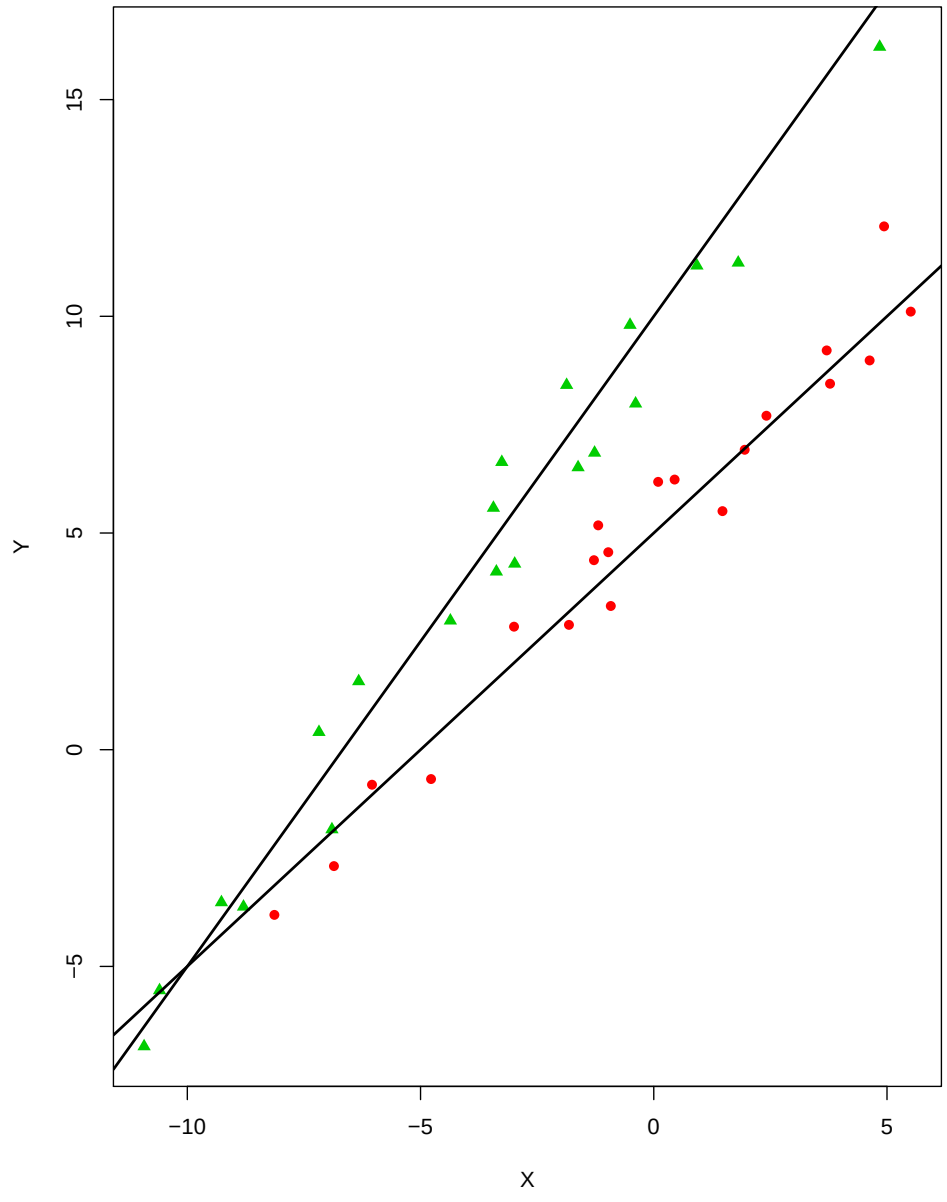
b_1 = change in intercept from placebo to treatment

b_2 = slope for the placebo group

b_3 = change in slope from placebo to treatment

Lines are no longer parallel meaning the difference between the groups changes with the value of X

$$\mu_y = 5 + 5I(treat) + X + \frac{1}{2}XI(treat)$$



Regression estimates

```
##           Estimate Std. Error
## Intercept 6.691252 0.36336549
## Slope     1.088475 0.07506545
```

```
##          Estimate Std. Error
## Placebo 5.205411 0.26884259
## TRT-PLB 3.639710 0.41333761
## Slope   1.252337 0.04707134
```

```
##                  Estimate Std. Error
## Int(Placebo) 5.1452515 0.20744453
## Int(TRT-PLB) 4.3576134 0.34779714
## Slp(Placebo) 1.0535716 0.05305477
## Slp(TRT-PLB) 0.3730443 0.07268315
```

Sample Size and Power calculations

α (or Significance) is the chance of accepting what is false. It does not depend on n, the sample size, but is chosen. Want it to be small and is commonly set at 0.05.

$1 - \beta$ (or Power) is the chance of accepting what is true. It depends on n and increases as n increases. To calculate Power, you need to specify each parameter in the model (based on previous studies, educated guesses or some other source). The more complicated the model the more parameters you need to specify.

Java Applets to calculate sample size and power
`http://homepage.stat.uiowa.edu/~rlenth/Power/`

G*Power for Mac or Windows
`http://www.gpower.hhu.de/en.html`

Questions?

Department of Statistics at UBC:

www.stat.ubc.ca/how-can-you-get-help-your-data

SOS Program - An hour of free consulting to UBC graduate students. Funded by the Provost and VP Research Office.

STAT 551 - Stat grad students taking this course offer free statistical advice. Fall semester every academic year.

Short Term Consulting Service - Advice from Stat grad students. Fee-for-service on small projects (less than 15 hours).

Hourly Projects - ASDa professional staff. Fee-for-service consulting.