ANOVA and regression

Vanessa McNealis

Computational and Data Systems Initiative
McGill University

vanessa.mcnealis@mail.mcgill.ca

January 24, 2022

Note: The following material is heavily based on the lecture notes prepared by Professor Erica Moodie for the course EPIB 607 given at McGill University.

ANOVA and Regression

Workshop content

ANOVA and regression

- Linear regression
- Least squares method
- Estimation of parameters in regression
- Interpretation of regression parameters
- Model-checking
- Analysis of Variance (ANOVA)
- Correlation

Expectations and learning outcomes

- It's a mixed audience; some participants have more experience with statistical methods.
- Those who are completely new to linear regression will learn the essential tools and be able to implement them in their own projects.
- Those with some background will review the topic in a formal manner and gain a deeper understanding of the tools.

Regression and correlation

- Are age and cholesterol *related* to each other?
- How can we measure the *strength* of such relationships?
- Can we predict the *average* cholesterol of people who are 50 years old?
- Can we predict a 50-year-old *individual's* cholesterol?

Regression and correlation

- Are age and cholesterol *related* to each other?
- How can we measure the *strength* of such relationships?
- Can we predict the *average* cholesterol of people who are 50 years old?
- Can we predict a 50-year-old *individual's* cholesterol?

Simple linear regression

We will investigate models relating two quantities x and y through equations of the form

$$y = ax + b$$

where a and b are constants (that is, a we will relate y and x with a straight line).

Note: in regression, variables x and y are not treated exchangeably - we regard y as being a function of x (e.g. cholesterol as a function of age).

- y is called the response, or the dependent variable (cholesterol)
- x is called the predictor, the covariate, the explanatory variable, or the independent variable (age)
- Calling x an independent variable is confusing and not good practice (although unfortunately common)

Simple linear regression

We will investigate models relating two quantities x and y through equations of the form

$$y = ax + b$$

where a and b are constants (that is, a we will relate y and x with a straight line).

Note: in regression, variables x and y are not treated exchangeably - we regard y as being a function of x (e.g. cholesterol as a function of age).

- y is called the response, or the dependent variable (cholesterol)
- x is called the predictor, the covariate, the explanatory variable, or the independent variable (age)
- Calling x an independent variable is confusing and not good practice (although unfortunately common)

Probabilistic models

In practice, the relationships we observe between two variables are not exact (i.e., not deterministic). A more useful model allows for the possibility that the system is not observed perfectly, that is, we do not observe (x, y) pairs that are always consistent with a simple functional relationship.

Probabilistic models

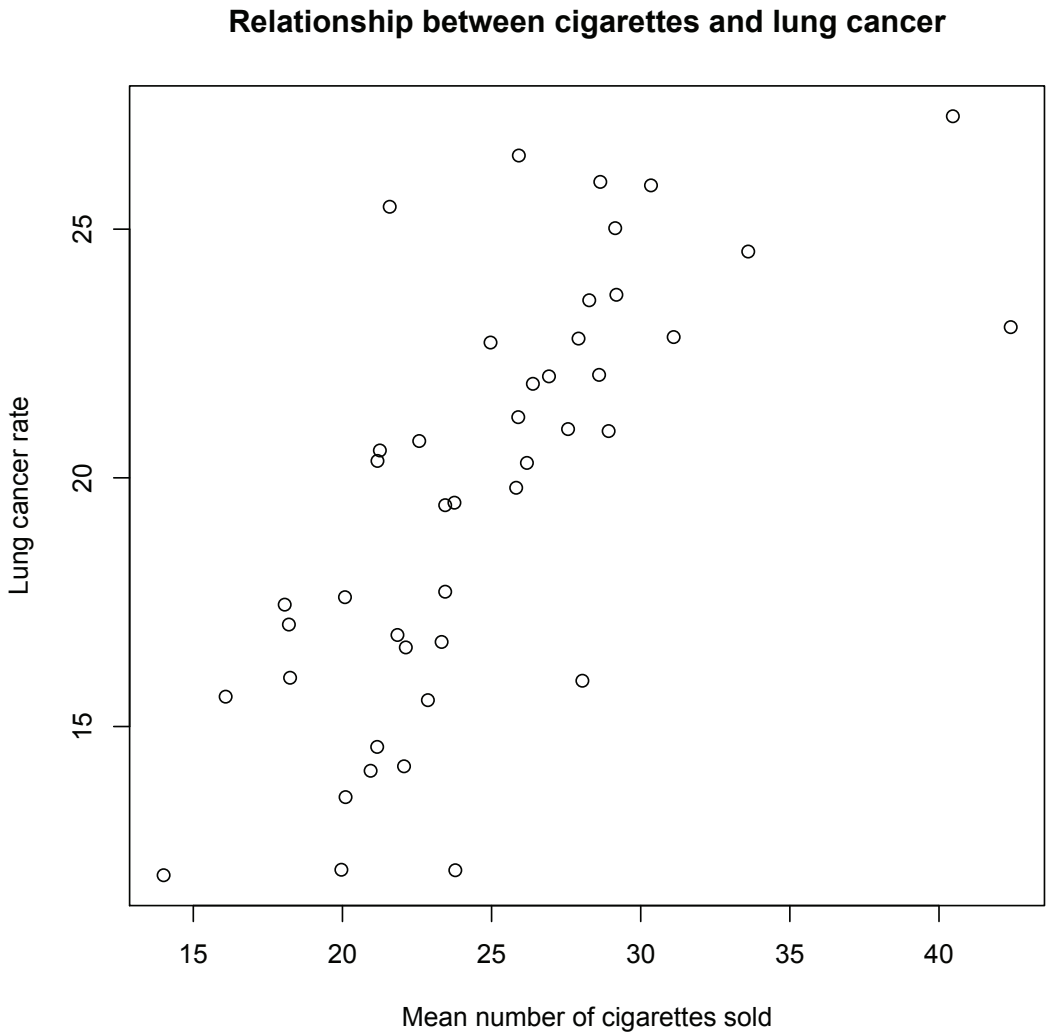
Let's consider one of the first studies (1960) to establish a formal link between smoking and cancer rates in USA. For 43 states in the USA, the following variables were measured:

1. STATE: state name
2. CIG: mean number of cigarettes sold per person in 1960 in the state.
3. BLAD: mortality rate (per 100,000) caused by bladder cancer in the state.
4. LUNG: mortality rate (per 100,000) caused by lung cancer in the state.
5. KID: mortality rate (per 100,000) caused by kidney cancer in the state.
6. LEUK: mortality rate (per 100,000) caused by leukemia in the state.

In this study, we want to establish the relationship between the number of cigarettes sold per person and the cancer rate. We'll only consider lung cancer for illustration purposes.

Probabilistic models

The graphical relationship between the number of cigarettes sold (x) and the rate of lung cancer (y) is given below:



Probabilistic models

To model such an "imperfect" relationship between x (cigarettes) and y (cancer rate), we use the following model:

$$y = ax + b + ERROR$$

where *ERROR* is a random term that is present due to imperfect observation of the system due to (i) measurement error or (ii) missing information.

We observe pairs $(x_i, y_i), i = 1, \ldots, n$. We model the variation in y as a function of x.

Note again that we do not treat x and y exchangeably. We assume that x is a fixed observed variable that is measured **without error**, whereas y is an observed variable that is measured **with random error**.

Probabilistic models

To model such an "imperfect" relationship between x (cigarettes) and y (cancer rate), we use the following model:

$$y = ax + b + ERROR$$

where *ERROR* is a random term that is present due to imperfect observation of the system due to (i) measurement error or (ii) missing information.

We observe pairs $(x_i, y_i), i = 1, \ldots, n$. We model the variation in y as a function of x.

Note again that we do not treat x and y exchangeably. We assume that x is a fixed observed variable that is measured **without error**, whereas y is an observed variable that is measured **with random error**.

A basic probabilistic model for linear regression

Using more typical notation, the model we study takes the form

$$y = \beta_0 + \beta_1 x + \epsilon$$

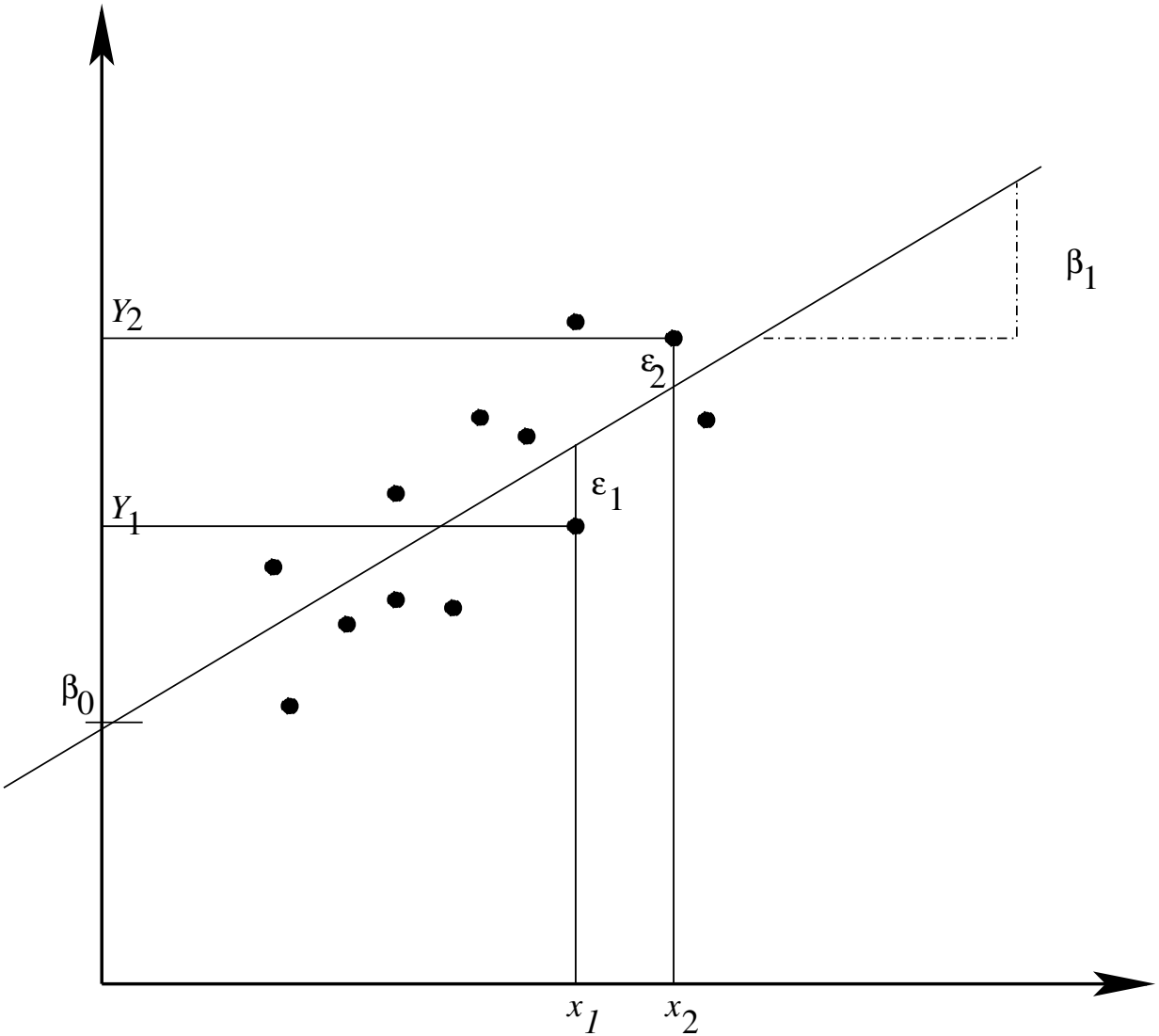
where ϵ is a random error term, a random variable with zero mean and finite variance ($E[\epsilon] = 0$, $Var[\epsilon] = \sigma^2$); it represents the error present in the measurement of y.

Terminology:

- β_0 - **Intercept** parameter
- β_1 - **Slope** parameter

A basic probabilistic model for linear regression

Graphically, these parameters are represented as the following:



A basic probabilistic model for linear regression

- $\beta_1 > 0$: increasing y with increasing x
- $\beta_1 < 0$: decreasing y with increasing x
- $\beta_1 = 0$: no *linear* relationship between x and y

Note:

$$E[Y|x] = \beta_0 + \beta_1 x$$

where $E[Y|x]$ is the expected value of Y for fixed value of x.

We use the notation

- Y - a random variable with a probability distribution
- y - a fixed value that the variable Y can take.

A basic probabilistic model for linear regression

- $\beta_1 > 0$: increasing y with increasing x
- $\beta_1 < 0$: decreasing y with increasing x
- $\beta_1 = 0$: no *linear* relationship between x and y

Note:

$$E[Y|x] = \beta_0 + \beta_1 x$$

where $E[Y|x]$ is the expected value of Y for fixed value of x.

We use the notation

- Y - a random variable with a probability distribution
- y - a fixed value that the variable Y can take.

A basic probabilistic model for linear regression

Fundamental Problem:
We only have the observed data $\{(x_i, y_i), i = 1, \ldots, n\}$, which do not fall *exactly* on a line. If we believe the straight-line model with error is correct, how do we find the estimates of the parameters β_0 and β_1?

Linear regression: Least-Squares fitting

We select the best values of β_0 and β_1 by minimizing the **error in fit**. For example, for the two data points (x_1, y_1) and (x_2, y_2), the errors in fit are:

$$\begin{aligned} e_1 &= y_1 - (\beta_0 + \beta_1 x_1) \\ e_2 &= y_2 - (\beta_0 + \beta_1 x_2) \end{aligned}$$

respectively.

But note that, potentially, $e_1 > 0$ and $e_2 < 0$ so there is a possibility that these fitting errors cancel each other out.

Therefore, it is possible that if we used the criterion of minimizing the sum of the errors to find estimates that we could have very big errors that sum to something near zero.

Linear regression: Least-Squares fitting

We select the best values of β_0 and β_1 by minimizing the **error in fit**. For example, for the two data points (x_1, y_1) and (x_2, y_2), the errors in fit are:

$$\begin{aligned} e_1 &= y_1 - (\beta_0 + \beta_1 x_1) \\ e_2 &= y_2 - (\beta_0 + \beta_1 x_2) \end{aligned}$$

respectively.

But note that, potentially, $e_1 > 0$ and $e_2 < 0$ so there is a possibility that these fitting errors cancel each other out.

Therefore, it is possible that if we used the criterion of minimizing the sum of the errors to find estimates that we could have very big errors that sum to something near zero.

Linear regression: Least-Squares fitting

Therefore we look at **squared** errors (which penalize large negative errors as much as large positive errors):

$$\begin{aligned} e_1^2 &= (y_1 - (\beta_0 + \beta_1 x_1))^2 \\ e_2^2 &= (y_2 - (\beta_0 + \beta_1 x_2))^2 \end{aligned}$$

Linear regression: Least-Squares fitting

For n data points, we obtain n squared errors

$$e_1^2, e_2^2, \ldots, e_n^2.$$

We select β_0 and β_1 as the values of the parameters that minimize the sum of the squared errors, $SSE = SSE(\beta_0, \beta_1)$, where

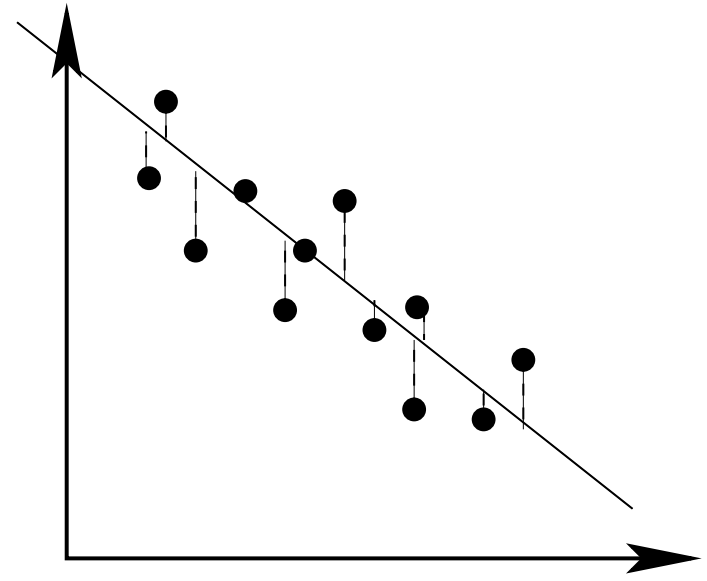
$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

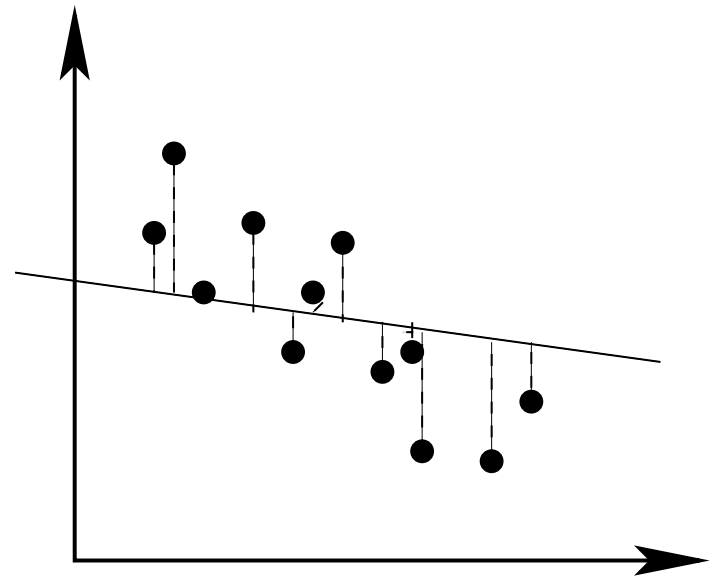
That is, we try to make the total misfit of the model (the squared error) as small as possible.

Minimization of $SSE(\beta_0, \beta_1)$ is achieved analytically.

Linear regression: Least-Squares fitting

Vanessa McNealis





Linear regression: Least-Squares fitting

Two routes: (i) calculus and (ii) geometric methods. It follows that the best parameters $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are given by

$$\widehat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

where

- Sum of Squares of x, SS_{xx}:

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

- Sum of Squares of x and y, SS_{xy}:

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Linear regression: Least-Squares fitting

$\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the **least-squares estimates**, and

$$y = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

is the **least-squares line of best fit**.

The **fitted values** are

$$\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \qquad i = 1, \ldots, n$$

and the **residuals** or **residual errors** are

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \qquad i = 1, \ldots, n.$$

Linear regression: Least-Squares fitting

$\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the **least-squares estimates**, and

$$y = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

is the **least-squares line of best fit**.

The **fitted values** are

$$\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \qquad i = 1, \ldots, n$$

and the **residuals** or **residual errors** are

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \qquad i = 1, \ldots, n.$$

Model assumptions for least-squares

To use least-squares for the probabilistic model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

we make the following assumptions

1. The expected error $E[\epsilon]$ is zero so that

$$E[Y|x] = \beta_0 + \beta_1 x$$

2. The variance of the error, $Var[\epsilon]$, is constant, finite, and does not depend on x.
3. The probability distribution of ϵ is Normal (a weaker assumption is that ϵ is symmetrically distributed).
4. The errors for two different measured responses are independent, i.e. the error ϵ_1 in measuring y_1 at x_1 is independent of the error ϵ_2 in measuring y_2 at x_2.

Model assumptions for least-squares

To use least-squares for the probabilistic model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

we make the following assumptions

1. The expected error $E[\epsilon]$ is zero so that

$$E[Y|x] = \beta_0 + \beta_1 x$$

2. The variance of the error, $Var[\epsilon]$, is constant, finite, and does not depend on x.
3. The probability distribution of ϵ is Normal (a weaker assumption is that ϵ is symmetrically distributed).
4. The errors for two different measured responses are independent, i.e. the error ϵ_1 in measuring y_1 at x_1 is independent of the error ϵ_2 in measuring y_2 at x_2.

Model assumptions for least-squares

To use least-squares for the probabilistic model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

we make the following assumptions

1. The expected error $E[\epsilon]$ is zero so that

$$E[Y|x] = \beta_0 + \beta_1 x$$

2. The variance of the error, $Var[\epsilon]$, is constant, finite, and does not depend on x.
3. The probability distribution of ϵ is Normal (a weaker assumption is that ϵ is symmetrically distributed).
4. The errors for two different measured responses are independent, i.e. the error ϵ_1 in measuring y_1 at x_1 is independent of the error ϵ_2 in measuring y_2 at x_2.

Model assumptions for least-squares

To use least-squares for the probabilistic model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

we make the following assumptions

1. The expected error $E[\epsilon]$ is zero so that

$$E[Y|x] = \beta_0 + \beta_1 x$$

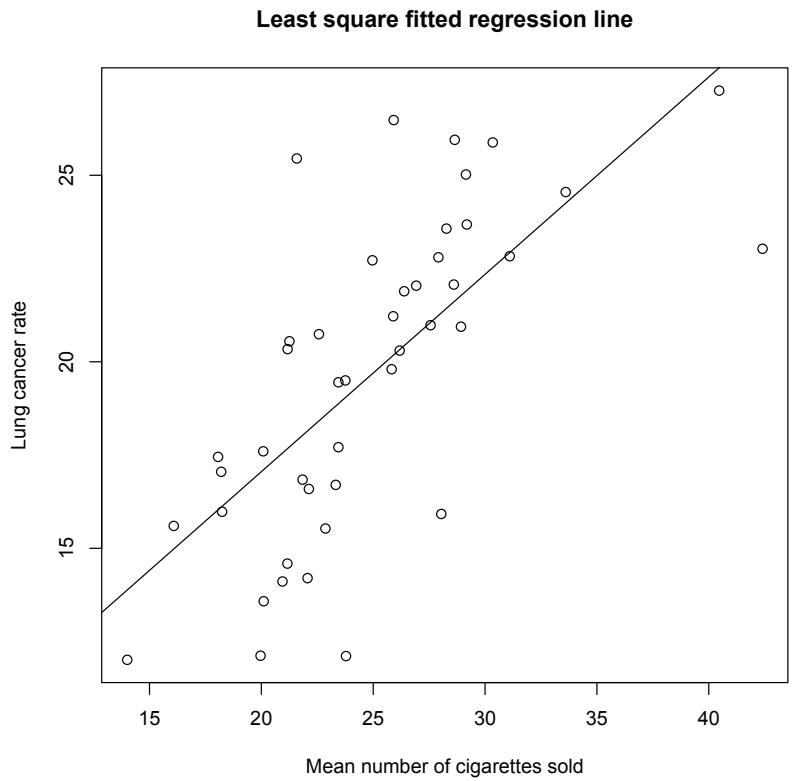
2. The variance of the error, $Var[\epsilon]$, is constant, finite, and does not depend on x.
3. The probability distribution of ϵ is Normal (a weaker assumption is that ϵ is symmetrically distributed).
4. The errors for two different measured responses are independent, i.e. the error ϵ_1 in measuring y_1 at x_1 is independent of the error ϵ_2 in measuring y_2 at x_2.

Linear regression

If we go back to the smoking dataset, the least-squares estimates are:

$$\hat{\beta}_1 = 0.5291, \qquad \hat{\beta}_0 = 6.4717.$$

The fitted regression line is shown below:



Linear regression parameter estimation for σ^2

Using the LS procedure, we can construct an estimate of the **error variance** or **residual error variance**.

Recall that

$$Var[\epsilon] = \sigma^2.$$

An estimate of σ^2 is

$$\widehat{\sigma}^2 = \frac{SSE(\widehat{\beta}_0, \widehat{\beta}_1)}{n-2},$$

where

$$SSE(\widehat{\beta}_0, \widehat{\beta}_1) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SS_{yy} - \widehat{\beta}_1 SS_{xy}$$

for $SS_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2$.

Linear regression parameter estimation for σ^2

Note that the denominator $n - 2$ is again a *degrees of freedom* parameter of the form

$$\text{TOTAL NUMBER OF DATA POINTS} - \text{NUMBER OF PARAMETERS ESTIMATED}$$

or $n - p$, where in simple linear regression with a single covariate x, we have $p = 2$ (β_0 and β_1).

Interpreting linear regression coefficients

How do we interpret the parameters of a linear regression?

Y, as always, is quantitative. Let's start with x a quantitative variable as well.

$$E[Y|x] = \beta_0 + \beta_1 x$$

- β_1 is the expected difference in response Y between two groups of individuals, where one group has covariate value x one unit greater than the other group.
- β_0 is the expected value of Y in a group of individuals with covariate value $x = 0$.

Interpreting linear regression coefficients

Less formally,

- β_1 is the expected change in Y associated with a 1-unit increase in x.
- β_0 is the expected value of Y in a group of individuals with $x = 0$.

This interpretation is correct, but always remember that these are **associative** models. Obtaining an estimate from a linear regression model does not imply that a causal effect has been identified.

Interpreting linear regression coefficients

Cautions:

- It may be the case that $x = 0$ is impossible or implausible (e.g. what if x is blood pressure?). If so, we should not "over-interpret" β_0.

- We should not try to make predictions outside of the observed range of the covariate. This is called *extrapolation*; the relationship between x and y may not be linear outside the observed range of the data (we simply have no data to judge this).

Interpreting linear regression coefficients

Cautions:

- It may be the case that $x = 0$ is impossible or implausible (e.g. what if x is blood pressure?). If so, we should not "over-interpret" β_0.

- We should not try to make predictions outside of the observed range of the covariate. This is called *extrapolation*; the relationship between x and y may not be linear outside the observed range of the data (we simply have no data to judge this).

Interpretation: when x is binary

Suppose the covariate that we wish to consider is binary, for example $x = 1$ if an individual in male and 0 otherwise.

Returning to our basic interpretation of the linear model, we have:

$$E[Y|x] = \beta_0 + \beta_1 x$$

- β_1 is the expected difference in response Y between two groups of individuals, where one group has covariate value x one unit greater than the other group.
- β_0 is the expected value of Y in a group of individuals with covariate value $x = 0$.

How could this model be extended to a factor variable that takes more than two values?

Estimation and testing for slope

In the model where

$$E[Y|x] = \beta_0 + \beta_1 x$$

it is of interest to test the hypothesis

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_A &: \beta_1 \neq 0 \end{aligned}$$

i.e. H_0 implies that there is no systematic linear contribution of x to the variation of y. That is, if we assume that x and y are linearly related, then $\beta_1 = 0$ implies that knowing x doesn't help us to predict Y (in a linear model).

Estimation and testing for slope

To test H_0 vs H_A we us the test statistic

$$t = \frac{\widehat{\beta}_1}{\widehat{se}(\widehat{\beta}_1)}$$

where $\widehat{se}(\widehat{\beta}_1)$ is the **estimated standard error** of $\widehat{\beta}_1$, computed as

$$\widehat{se}(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{\sqrt{SS_{xx}}}$$

where $\widehat{\sigma}^2$ is the estimator of σ^2 defined previously.

If H_0 is true, and $\beta_1 = 0$, then

$$t = \frac{\widehat{\beta}_1}{\widehat{\sigma}/\sqrt{SS_{xx}}} \sim t(n-2)$$

so we can carry out a significance test at level α in the usual way (use a p-value, or construct the rejection region). One-sided tests are also possible.

Estimation and testing for slope

To test H_0 vs H_A we us the test statistic

$$t = \frac{\widehat{\beta}_1}{\widehat{se}(\widehat{\beta}_1)}$$

where $\widehat{se}(\widehat{\beta}_1)$ is the **estimated standard error** of $\widehat{\beta}_1$, computed as

$$\widehat{se}(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{\sqrt{SS_{xx}}}$$

where $\widehat{\sigma}^2$ is the estimator of σ^2 defined previously.

If H_0 is true, and $\beta_1 = 0$, then

$$t = \frac{\widehat{\beta}_1}{\widehat{\sigma}/\sqrt{SS_{xx}}} \sim t(n-2)$$

so we can carry out a significance test at level α in the usual way (use a p-value, or construct the rejection region). One-sided tests are also possible.

Estimation and testing for slope

Note: To test

$$\begin{aligned} H_0 &: \beta_1 = b \\ H_A &: \beta_1 \neq b \end{aligned}$$

for any b, the test statistic is

$$t = \frac{\widehat{\beta}_1 - b}{\widehat{\sigma}/\sqrt{SS_{xx}}}$$

(for example, $b = 1$ may be of interest). If H_0 is true

$$t \sim t(n-2)$$

Confidence Interval

A $100(1-\alpha)\%$ confidence interval for β_1 is

$$\widehat{\beta}_1 \pm t^*_{\alpha/2}(n-2) \times \widehat{se}(\widehat{\beta}_1)$$

where $t^*_{\alpha/2}(n-2)$ is the $\alpha/2$ quantile of $t(n-2)$ distribution.

Note: we could perform a similar analysis for β_0, but this is generally of less interest. If we want to perform such a test, we need the estimated standard error of $\widehat{\beta}_0$. It can be shown that

$$\widehat{se}(\widehat{\beta}_0) = \sqrt{\frac{\widehat{\sigma}^2}{n}\left(1 + \frac{n\overline{x}^2}{SS_{xx}}\right)}.$$

Confidence Interval

A $100(1-\alpha)\%$ confidence interval for β_1 is

$$\widehat{\beta}_1 \pm t^*_{\alpha/2}(n-2) \times \widehat{se}(\widehat{\beta}_1)$$

where $t^*_{\alpha/2}(n-2)$ is the $\alpha/2$ quantile of $t(n-2)$ distribution.

Note: we could perform a similar analysis for β_0, but this is generally of less interest. If we want to perform such a test, we need the estimated standard error of $\widehat{\beta}_0$. It can be shown that

$$\widehat{se}(\widehat{\beta}_0) = \sqrt{\frac{\widehat{\sigma}^2}{n}\left(1 + \frac{n\bar{x}^2}{SS_{xx}}\right)}.$$

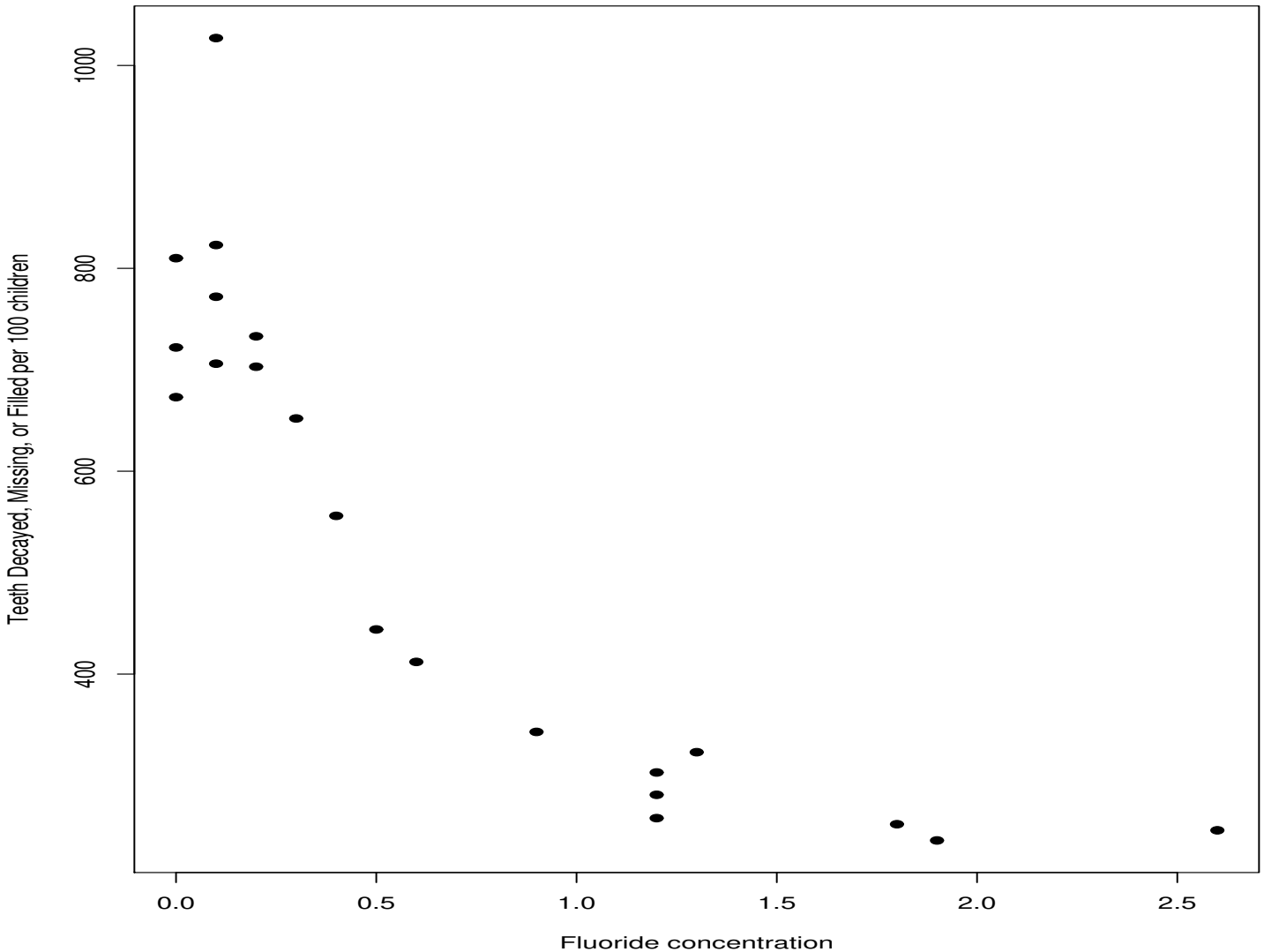
Example

The following data show the number of teeth that are decayed, missing, or filled as a rate per 100 children (DMF) in 21 communities, along with the fluoride concentration (Fconc) of the public water supply in those communities:

Community	DMF	Fconc		Community	DMF	Fconc
1	236	1.9		12	652	0.3
2	246	2.6		13	673	0.0
3	252	1.8		14	703	0.2
4	258	1.2		15	706	0.1
5	281	1.2		16	722	0.0
6	303	1.2		17	733	0.2
7	323	1.3		18	772	0.1
8	343	0.9		19	810	0.0
9	412	0.6		20	823	0.1
10	444	0.5		21	1027	0.1
11	556	0.4				

Example

Vanessa McNealis



Example

1. Calculate the regression line of DMF teeth on fluoride concentration.
2. Calculate the estimate of the residual standard error, σ.
3. Calculate a confidence interval for the slope, β_1, found in (1). What does that lead you to conclude?

Example

1. Calculate the regression line of DMF teeth on fluoride concentration.

$$\begin{aligned}\widehat{\beta}_1 &= \frac{SS_{xy}}{SS_{xx}} \\ &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} \\ &=\end{aligned}$$

$$\begin{aligned}\widehat{\beta}_0 &= \overline{y}-\widehat{\beta}_1\overline{x} \\ &=\end{aligned}$$

Therefore, the regression line is DMF = + Fconc.

Example

1. Calculate the regression line of DMF teeth on fluoride concentration.

$$\begin{aligned}\widehat{\beta}_1 &= \frac{SS_{xy}}{SS_{xx}} \\ &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} \\ &=\end{aligned}$$

$$\begin{aligned}\widehat{\beta}_0 &= \overline{y}-\widehat{\beta}_1\overline{x} \\ &=\end{aligned}$$

Therefore, the regression line is DMF = + Fconc.

Example

2. Calculate the estimate of the residual standard error, σ.

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{SSE(\widehat{\beta}_0, \widehat{\beta}_1)}{n-2} \\ &= \frac{SS_{yy} - \widehat{\beta}_1 SS_{xy}}{n-2} \\ &= \end{aligned}$$

So then $\widehat{\sigma} =$.

Example

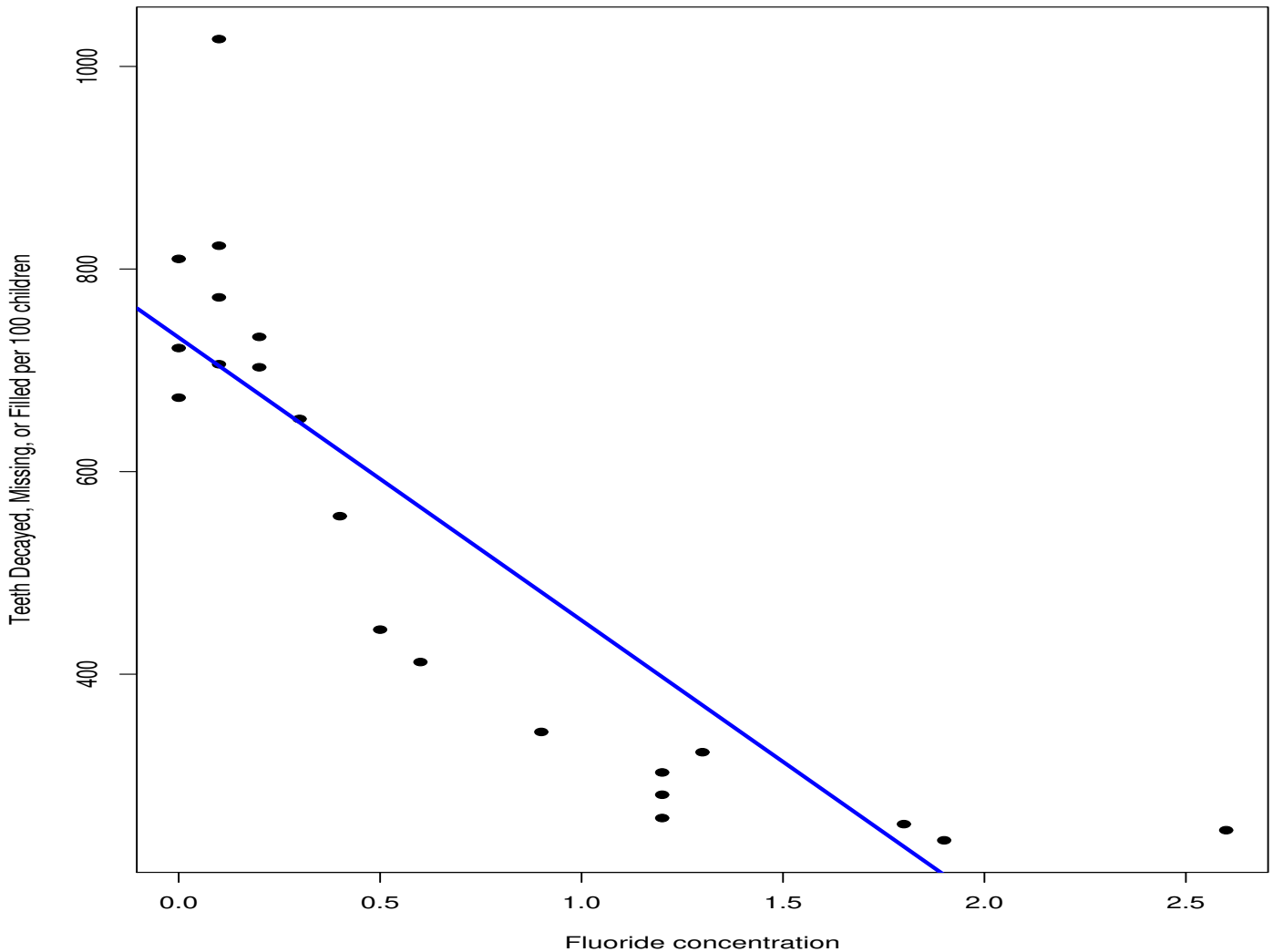
3. Calculate a confidence interval for the slope, β_1. What do you conclude?

$$\widehat{\beta}_1 \pm t^*_{(0.025,n-2)} \frac{\widehat{\sigma}}{\sqrt{SS_{xx}}} = \\ =$$

which gives a 95% confidence interval for the slope of (,). We conclude...

Example

...or do we? We'll soon see how we can use estimated residuals to check the fit of our model.



Example: R code

```
> dmf <-c(236,246,252,258,281,303,323,343,...)
> fconc <- c(1.9,2.6,1.8,1.2,1.2,1.2,1.3,0.9,...)
>
> n <- length(dmf)
>
> ## SS.yy:
> SS.dmf <- sum( (dmf - mean(dmf))^ 2 )
> ## SS.xx:
> SS.fconc <- sum( (fconc - mean(fconc))^ 2 )
>
> ## SS.xy:
> SS.dmf.fconc <-
   sum( (dmf - mean(dmf))*(fconc - mean(fconc)) )
```

Example: R code

```
> ## beta1.hat = SS.xy/SS.xx:
> beta1.hat <- SS.dmf.fconc/SS.fconc
> beta1.hat
[1] -279.1996
>
> beta0.hat <- mean(dmf) -
                 beta1.hat * mean(fconc)
> beta0.hat
[1] 732.3445
>
> sigma.hat <- sqrt( (SS.dmf -
            beta1.hat * SS.dmf.fconc)/(n-2) )
> sigma.hat
[1] 127.3215
```

Example: R code

```
> dmf.model <- lm(dmf~fconc)
> summary(dmf.model)

Call:
lm(formula = dmf ~ fconc)

Residuals:
     Min       1Q   Median       3Q      Max
-152.825  -94.305    1.575   56.495  322.575

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   732.34      38.55  18.996  8.1e-14 ***
fconc        -279.20      38.18  -7.312  6.2e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 127.3 on 19 degrees of freedom
Multiple R-Squared: 0.7378,     Adjusted R-squared: 0.724
F-statistic: 53.47 on 1 and 19 DF,  p-value: 6.199e-07
```

Prediction: averages and individuals

How can we predict the response for an individual with a particular value of x, or the average response at a particular value of x?

We use the fitted value as our estimate for both:

$$\hat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x.$$

It is "easier" to estimate an average response than an individual's response, so the standard error of the estimates are different.

Prediction: averages and individuals

How can we predict the response for an individual with a particular value of x, or the average response at a particular value of x?

We use the fitted value as our estimate for both:

$$\hat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x.$$

It is "easier" to estimate an average response than an individual's response, so the standard error of the estimates are different.

Prediction: averages and individuals

How can we predict the response for an individual with a particular value of x, or the average response at a particular value of x?

We use the fitted value as our estimate for both:

$$\hat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x.$$

It is "easier" to estimate an average response than an individual's response, so the standard error of the estimates are different.

Prediction: averages and individuals

Thus, the predicted or estimated *average* response for a particular value of x, say x^*, is

$$\hat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x^*,$$

and the estimated standard error of the prediction is

$$\widehat{SE}(\text{predicted } \textbf{mean} \text{ at } x) : \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{SS_{xx}}}.$$

Prediction: averages and individuals

The predicted or estimated response for an *individual* at particular value of x, say x^*, is

$$\hat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x^*,$$

and the estimated standard error of the prediction is

$$\widehat{SE}(\text{predicted } \textbf{individual} \text{ at } x) : \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{SS_{xx}}}.$$

Example, continued

4. Give a prediction for the average rate of DMF teeth for a community with a fluoride concentration of 1.8ppm.
5. Find a 95% confidence interval for your response to (4).

Example

4. Give a prediction for the average rate of DMF teeth for a community with a fluoride concentration of 1.8ppm.

$$\begin{aligned} \hat{y} &= \widehat{\beta}_0 + \widehat{\beta}_1 x^* \\ &= \\ &= \end{aligned}$$

Example

5. Find a 95% confidence interval for your response to (4).

$$\hat{y} \pm t^*_{(0.025,n-2)} \widehat{SE}(\text{predicted } \textbf{mean} \text{ at } x)$$

$$=$$

$$=$$

Example

5. Find a 95% confidence interval for your response to (4).

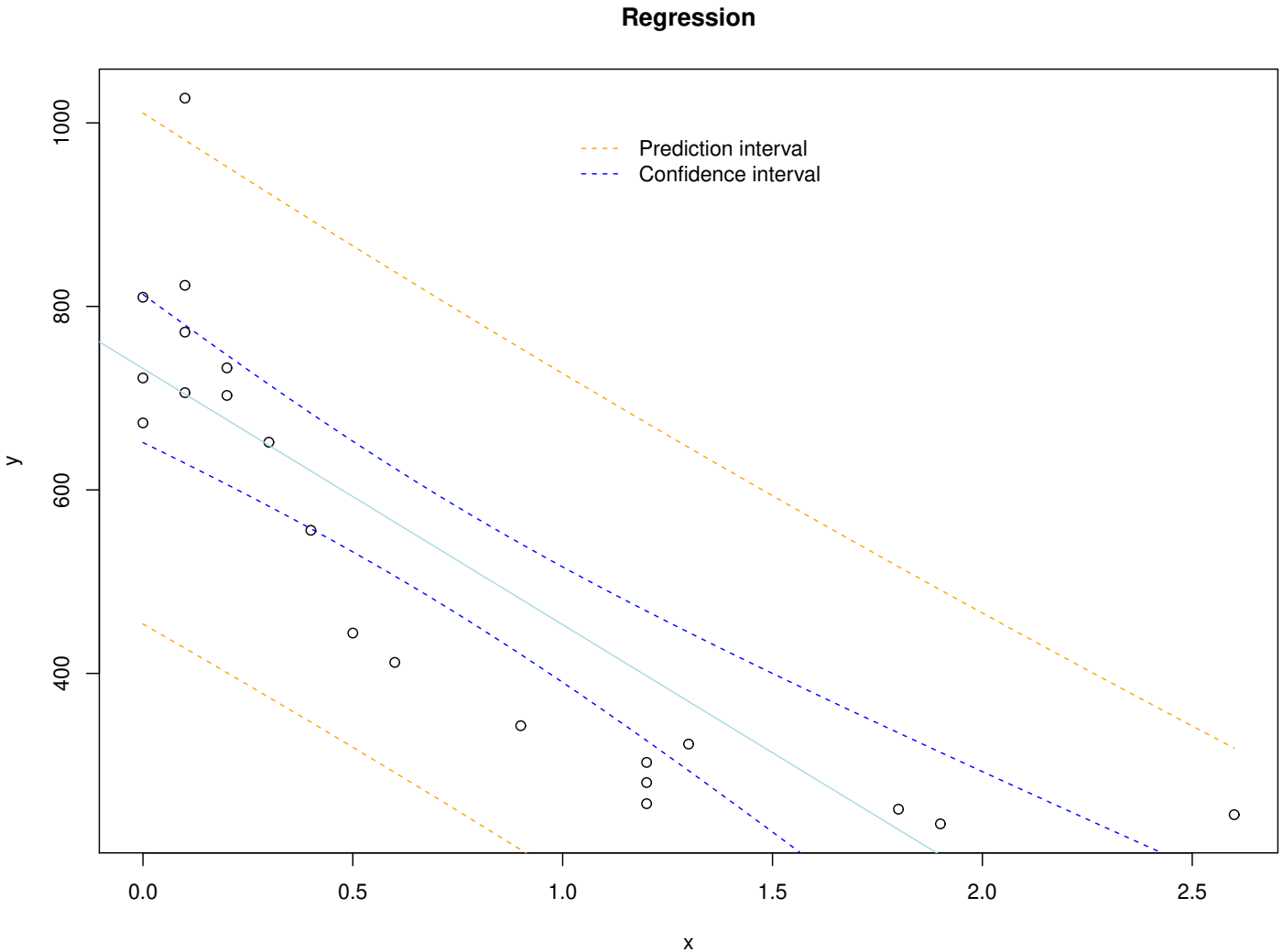
$$\hat{y} \pm t^*_{(0.025,n-2)} \widehat{SE}(\text{predicted } \textbf{mean} \text{ at } x)$$

$$=$$

$$=$$

Example

Vanessa McNealis



Model-checking

Recall that our model assumes:

1. The expected error $E[\epsilon]$ is zero so that

$$E[Y|x] = \beta_0 + \beta_1 x$$

2. The variance of the error, $Var[\epsilon]$, is constant, finite, and does not depend on x.
3. The probability distribution of ϵ is a Normal distribution, centred about zero.
4. The errors for two different measured responses are independent.

How can we verify whether these assumptions are met?

Model-checking

1. The expected error $E[\epsilon]$ is zero.

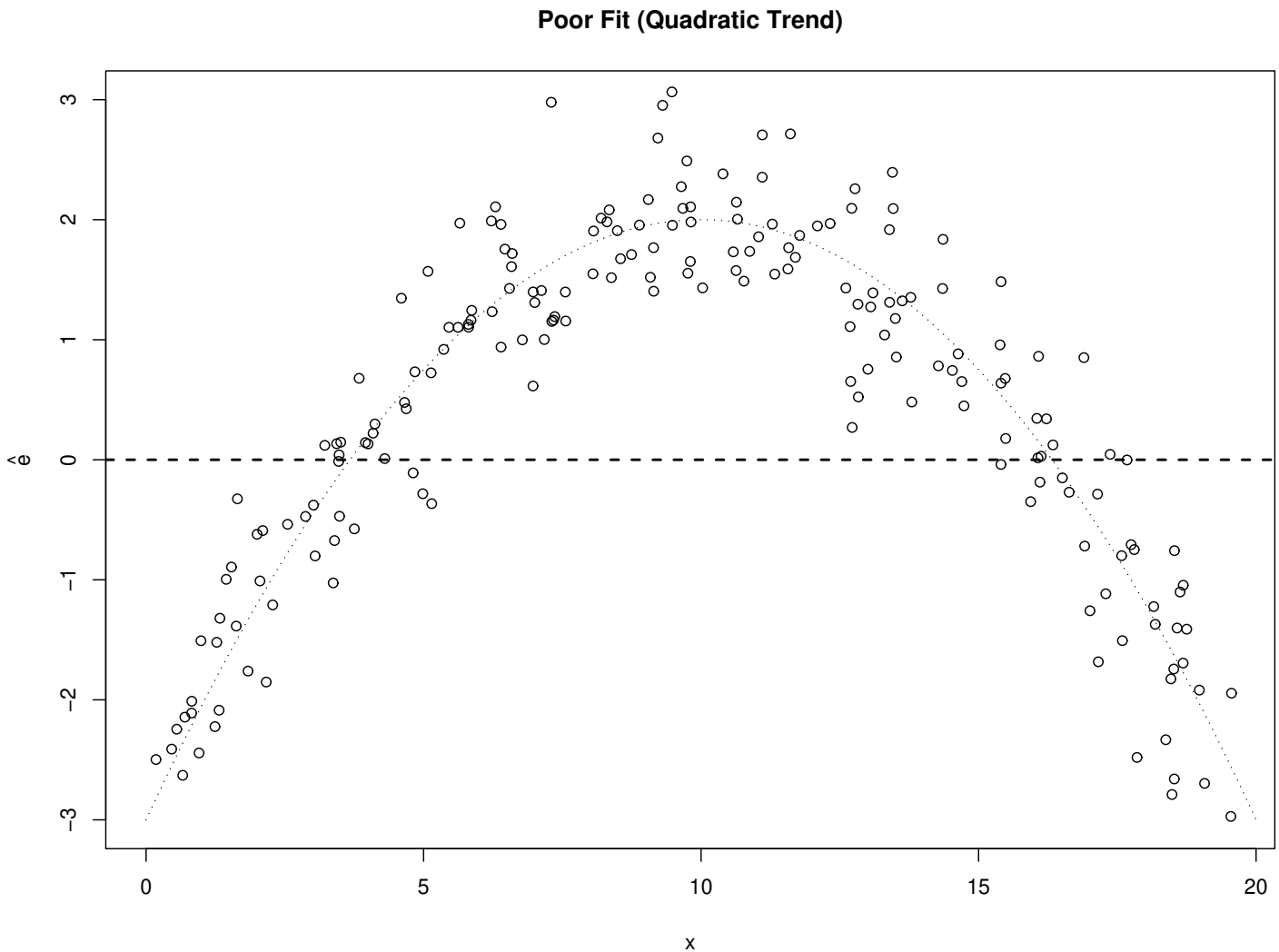
- The least squares fitting procedure ensures that this is the case.
- For the sake of argument, let's suppose that $E[Y|x] = a + bx$ and $E[\epsilon] = c$. Then taking $\beta_0 = a + c$ and $\beta_1 = b$ gives

$$E[Y|x] = \beta_0 + \beta_1 x$$

- That is, any non-zero mean of the errors would simply be absorbed into the intercept term without affecting the slope (which is typically what we are interested in).

Model-checking

We can also check whether a **linear** form is appropriate by plotting $\hat{e}_i$ against x:



Model-checking

2. The variance of the error, $Var[\epsilon]$, is constant and does not depend on x.

- We check this by plotting the **residuals**,

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

  (or estimated errors) against x.
- If we see a "cloud," we are satisfied that the assumption was met
- If we see a funnel-shape, we believe that the assumption was not met and consider a transformation of the data (e.g., log-transformation)

Model-checking

2. The variance of the error, $Var[\epsilon]$, is constant and does not depend on x.

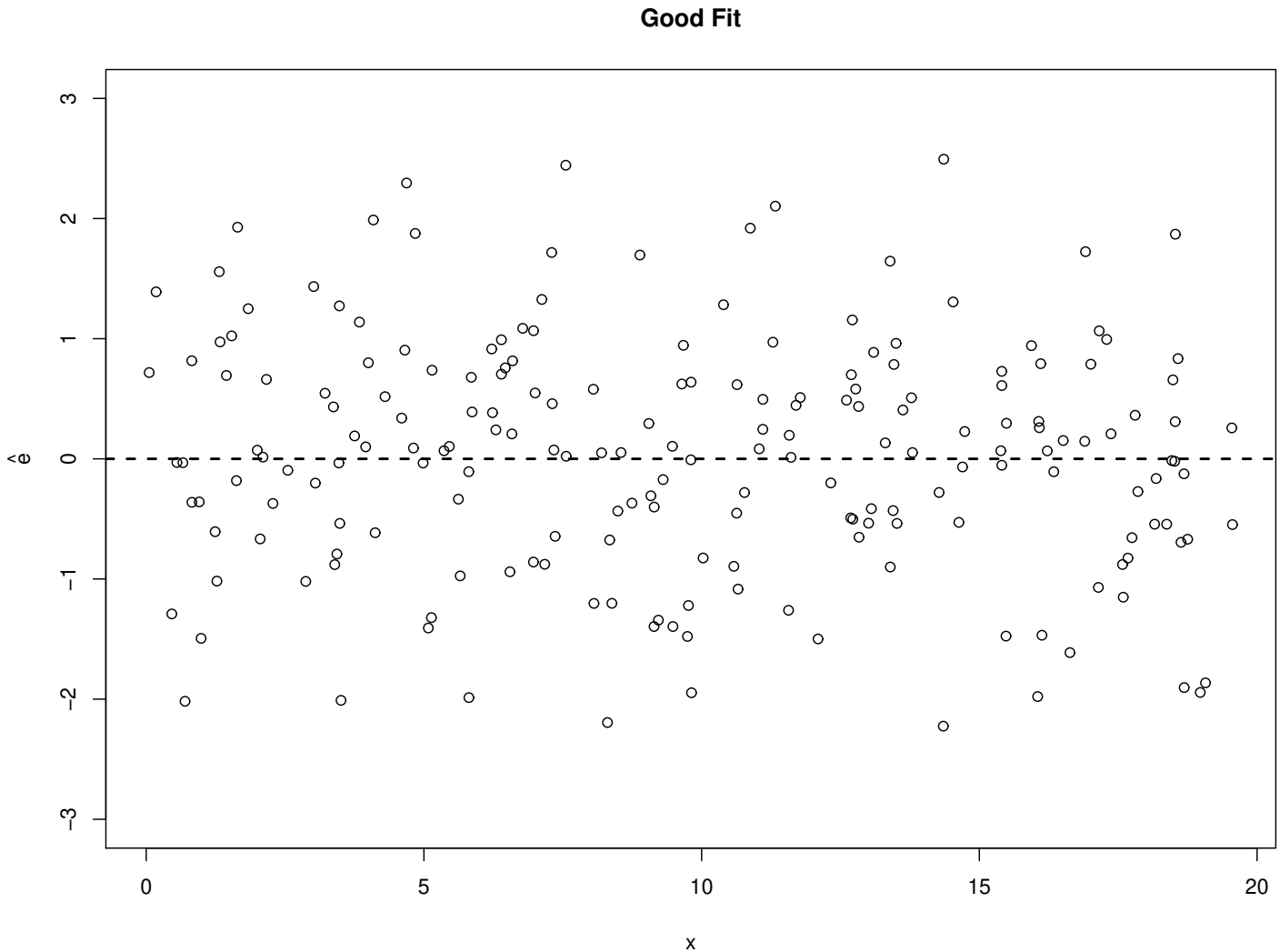
- We check this by plotting the **residuals**,

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

  (or estimated errors) against x.
- If we see a "cloud," we are satisfied that the assumption was met
- If we see a funnel-shape, we believe that the assumption was not met and consider a transformation of the data (e.g., log-transformation)

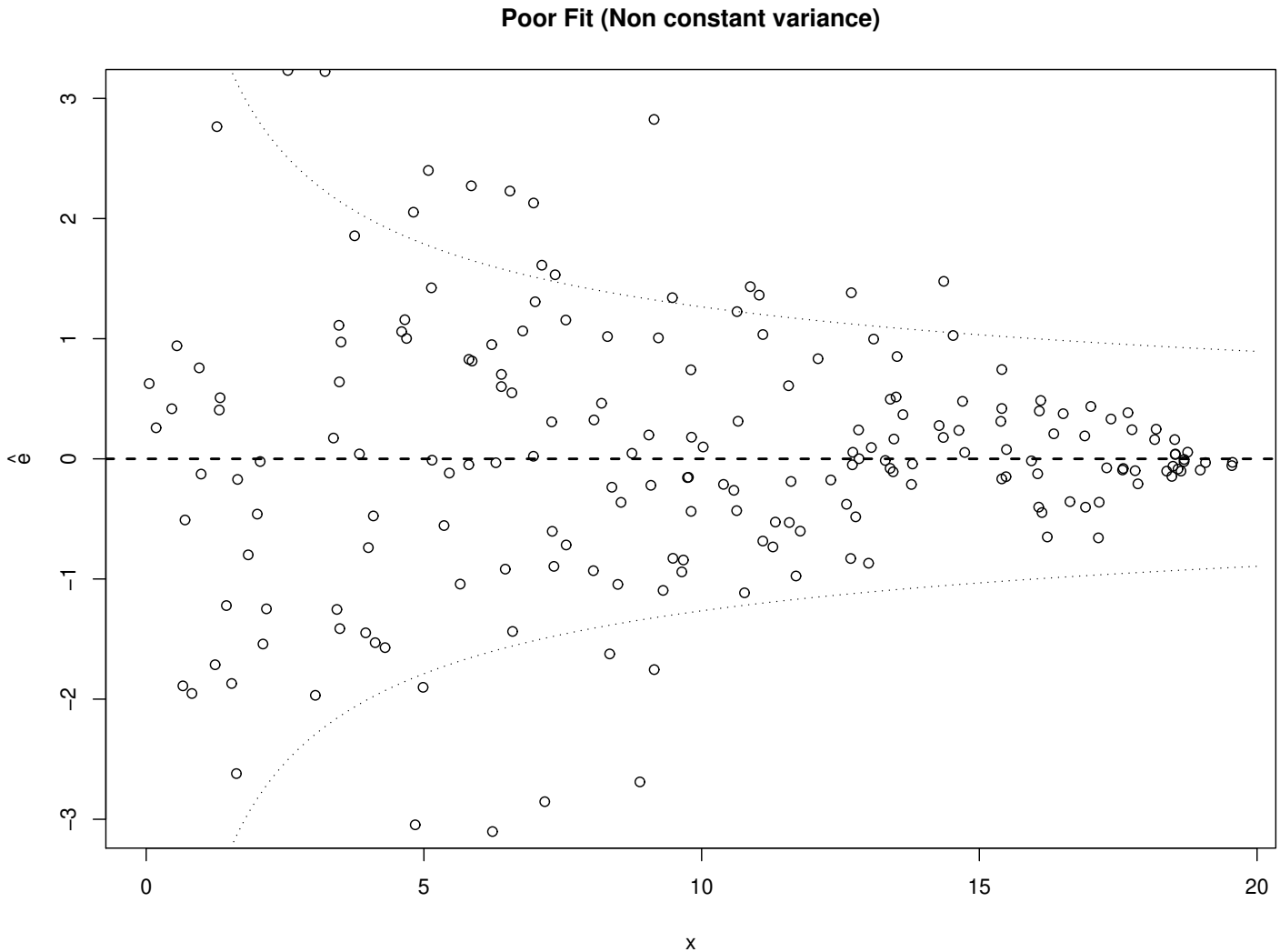
Model-checking

Vanessa McNealis



Model-checking

Vanessa McNealis



Model-checking

3. The probability distribution of ϵ is a Normal distribution centred about zero.

- This is most easily checked with a simple histogram, and an "eye-check" for symmetry/approximate Normality.
- We can also sometimes spot asymmetry in plots of the residuals against x.

Model-checking

4. The errors for two different measured responses are independent.

- Generally, we assume that this is true if our study is well-designed (random sampling!)
- If we know the order in which individuals entered into the study, we can look for patterns in the time-ordered residuals.

Model-checking

- It is also a good idea to plot the residuals against any other variables, $x2$, $x3$, etc that are available. If any trend is observed, this indicates that the covariate may be required in the linear regression model.

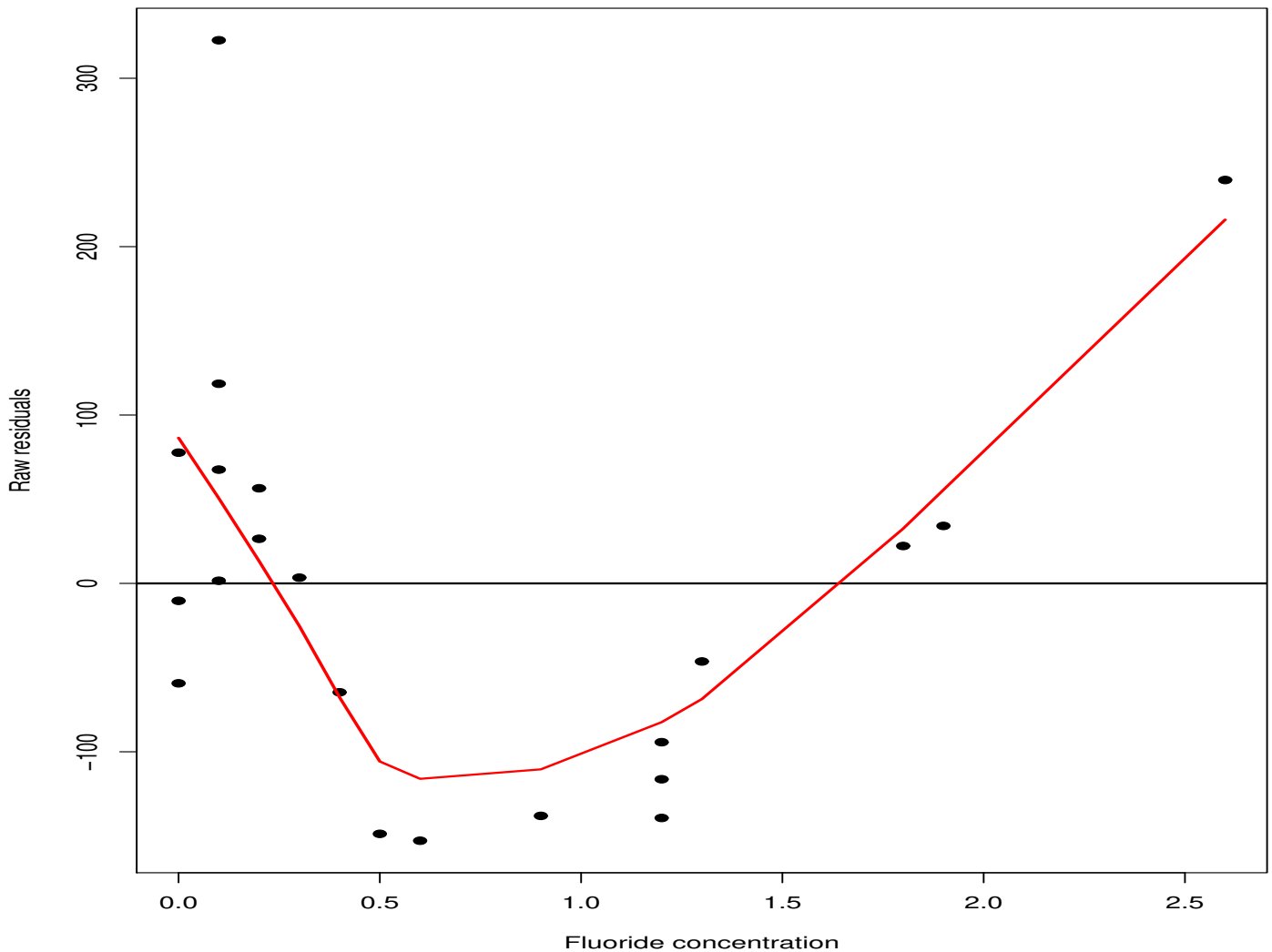
- Finally, plotting the residuals against the fitted values $\hat{y}$ may reveal trends in the variability (e.g., does the spread/variability of the residuals increase as $\hat{y}$ increases?). If so, a transformation of the response may be needed to stabilize the variance.

Model-checking

- It is also a good idea to plot the residuals against any other variables, $x2$, $x3$, etc that are available. If any trend is observed, this indicates that the covariate may be required in the linear regression model.

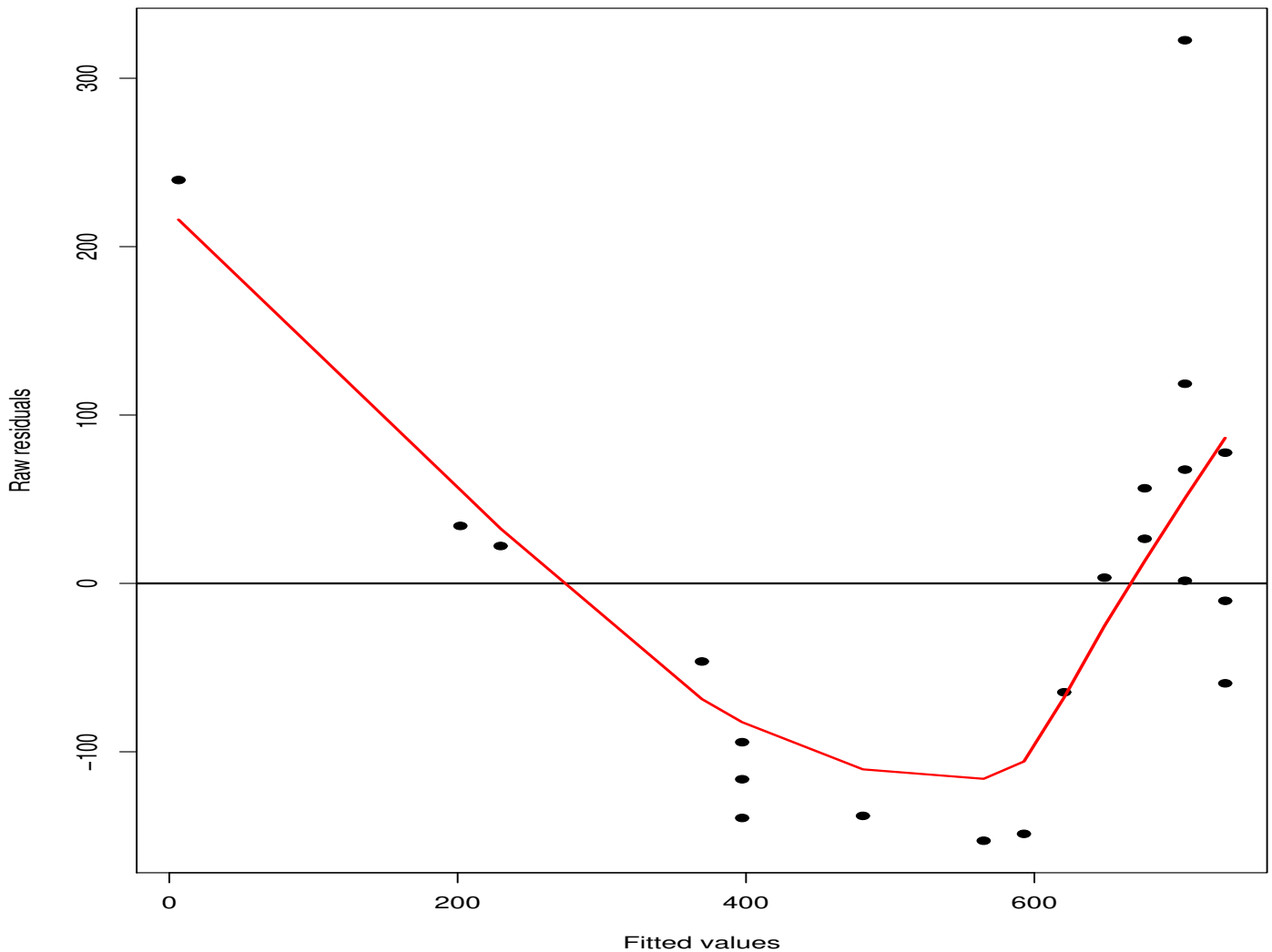
- Finally, plotting the residuals against the fitted values $\hat{y}$ may reveal trends in the variability (e.g., does the spread/variability of the residuals increase as $\hat{y}$ increases?). If so, a transformation of the response may be needed to stabilize the variance.

Model-checking: recall the DMF example...



Model-checking: recall the DMF example...

Vanessa McNealis



Model-checking: recall the DMF example...

```
> fitted.dmf <- beta0.hat + beta1.hat*fconc
> raw.resid <- dmf - fitted.dmf
>
> plot(fconc,raw.resid,pch=19,xlab="Fluoride
          concentration",ylab="Raw residuals")
> abline(0,0,lwd=2)
> lines(lowess(fconc,raw.resid),col=2,lwd=2)
>
> plot(fitted.dmf,raw.resid,pch=19,xlab="Fitted
                values",ylab="Raw residuals")
> abline(0,0,lwd=2)
> lines(lowess(fitted.dmf,raw.resid),col=2,lwd=2)
>
> hist(raw.resid,nclass = 8,xlab="Raw residuals")
> boxplot(raw.resid,ylab="Raw residuals")
> abline(0,0,lwd=2,col=4)
```

Model-checking: recall the DMF example...

Conclusions based on our model-diagnostics:

- Pattern suggests that the relationship between DMF and fluoride concentration might be better modelled with a quadratic relationship rather than a linear relationship.
- The linear regression was still useful in that it allowed us to examine the data and capture the overall negative association.
- However our confidence intervals are suspect since the distribution of the errors is not symmetric about 0 over the range of the covariate, and anyhow report confidence on a statistic that is not particularly useful.

Model-checking: recall the DMF example...

Conclusions based on our model-diagnostics:

- Pattern suggests that the relationship between DMF and fluoride concentration might be better modelled with a quadratic relationship rather than a linear relationship.
- The linear regression was still useful in that it allowed us to examine the data and capture the overall negative association.
- However our confidence intervals are suspect since the distribution of the errors is not symmetric about 0 over the range of the covariate, and anyhow report confidence on a statistic that is not particularly useful.

Model-checking: recall the DMF example...

Conclusions based on our model-diagnostics:

- Pattern suggests that the relationship between DMF and fluoride concentration might be better modelled with a quadratic relationship rather than a linear relationship.
- The linear regression was still useful in that it allowed us to examine the data and capture the overall negative association.
- However our confidence intervals are suspect since the distribution of the errors is not symmetric about 0 over the range of the covariate, and anyhow report confidence on a statistic that is not particularly useful.

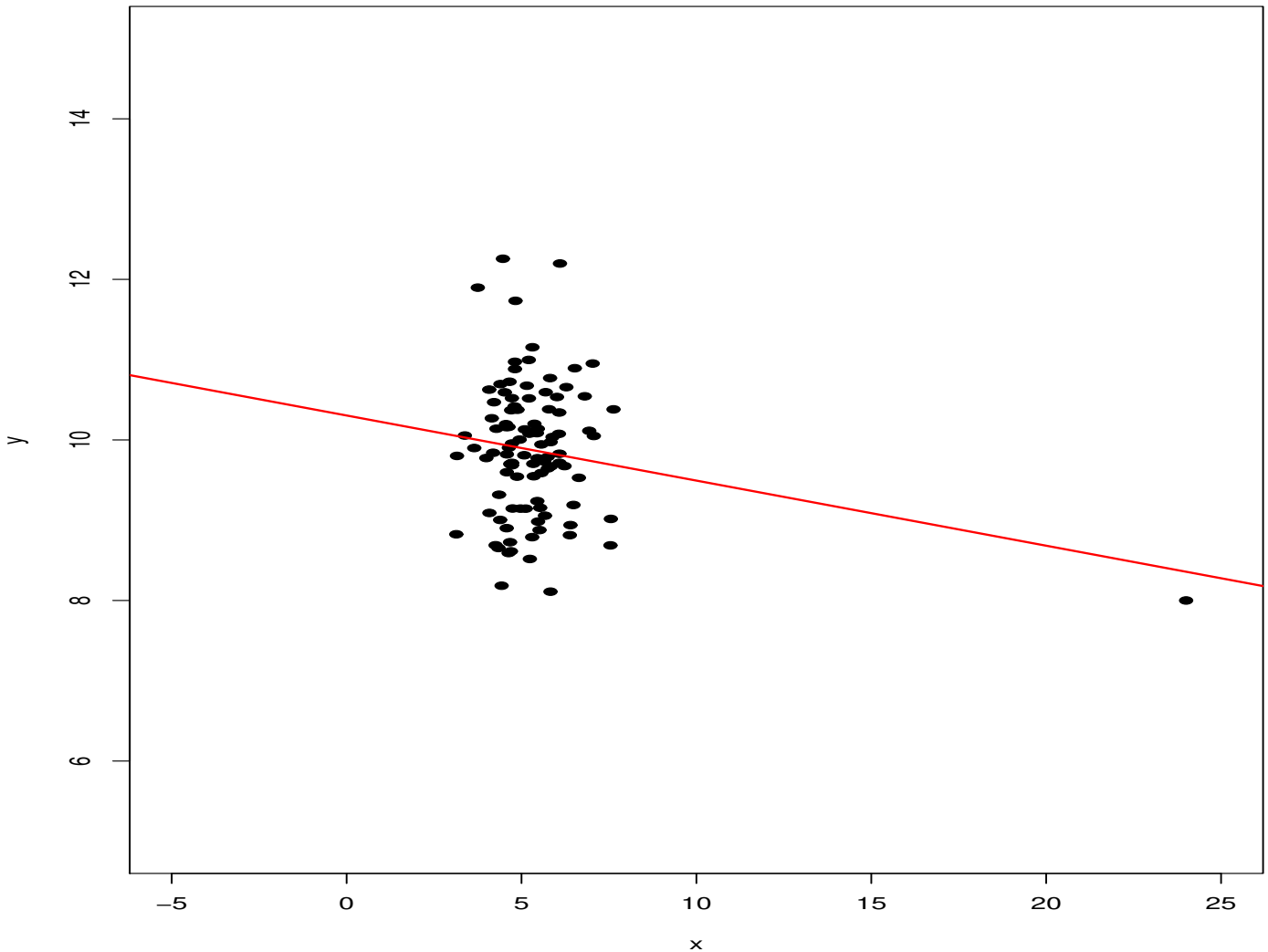
Model-checking: outliers

When examining scatter-plots or plots of the residuals, also look for outlying points.

- Points that are univariate outliers may not have much influence on the co-efficient estimates, which is why we need to examine the scatter plots to see whether points are outliers for the linear model that we have assumed.
- If one or two observations appear to be driving the relationship, you may want to transform the data to reduce the influence of these outlying points.

Model-checking: outliers

Vanessa McNealis

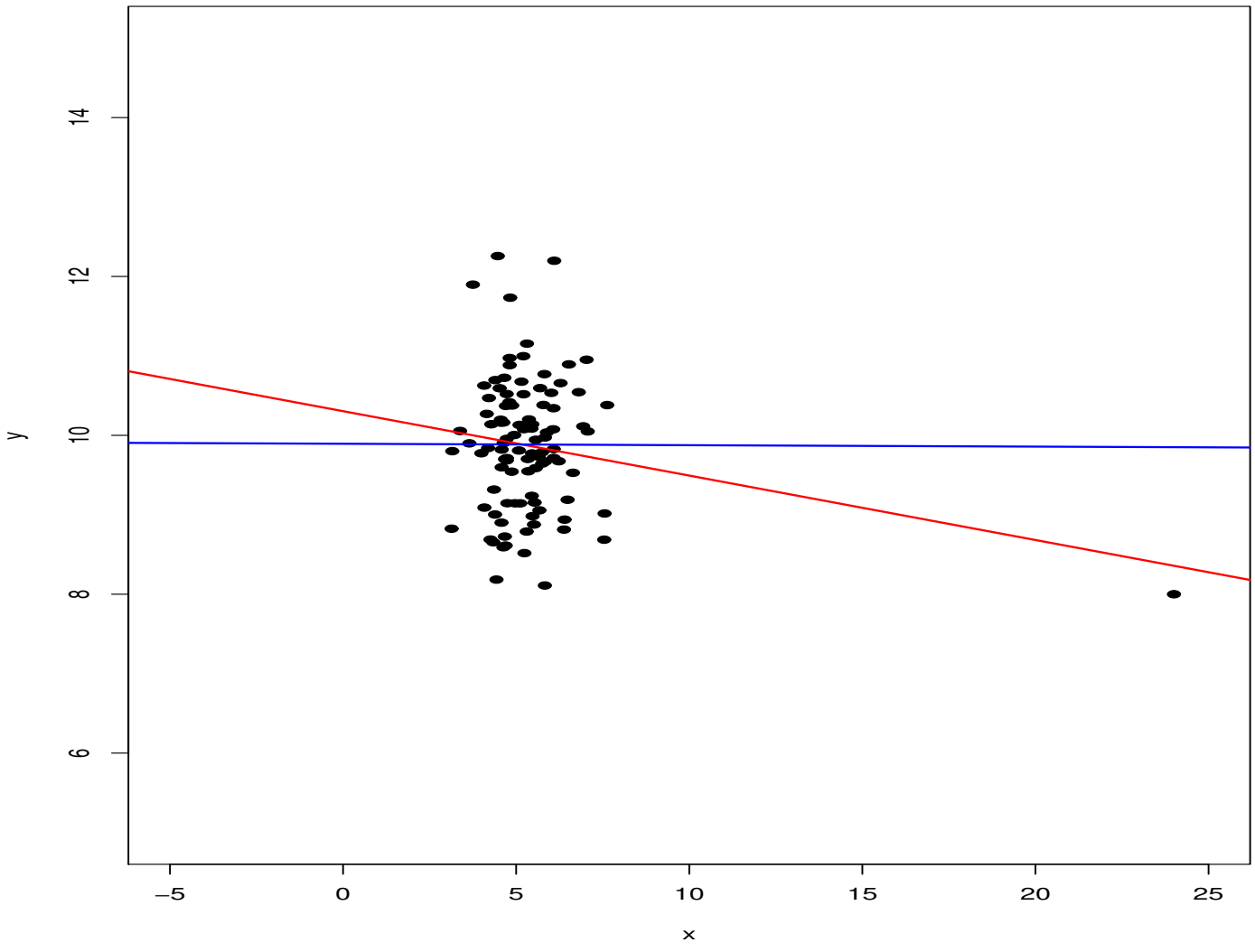


Model-checking: outliers


```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.76089 -0.62666 -0.03837  0.51309  2.38810

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.30429    0.22990  44.821   <2e-16 ***
x           -0.08110    0.03967  -2.045   0.0436 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8263 on 98 degrees of freedom
Multiple R-Squared: 0.04091,    Adjusted R-squared: 0.03112
F-statistic:  4.18 on 1 and 98 DF,  p-value: 0.04359
```

Model-checking: outliers

Vanessa McNealis



Model-checking: outliers

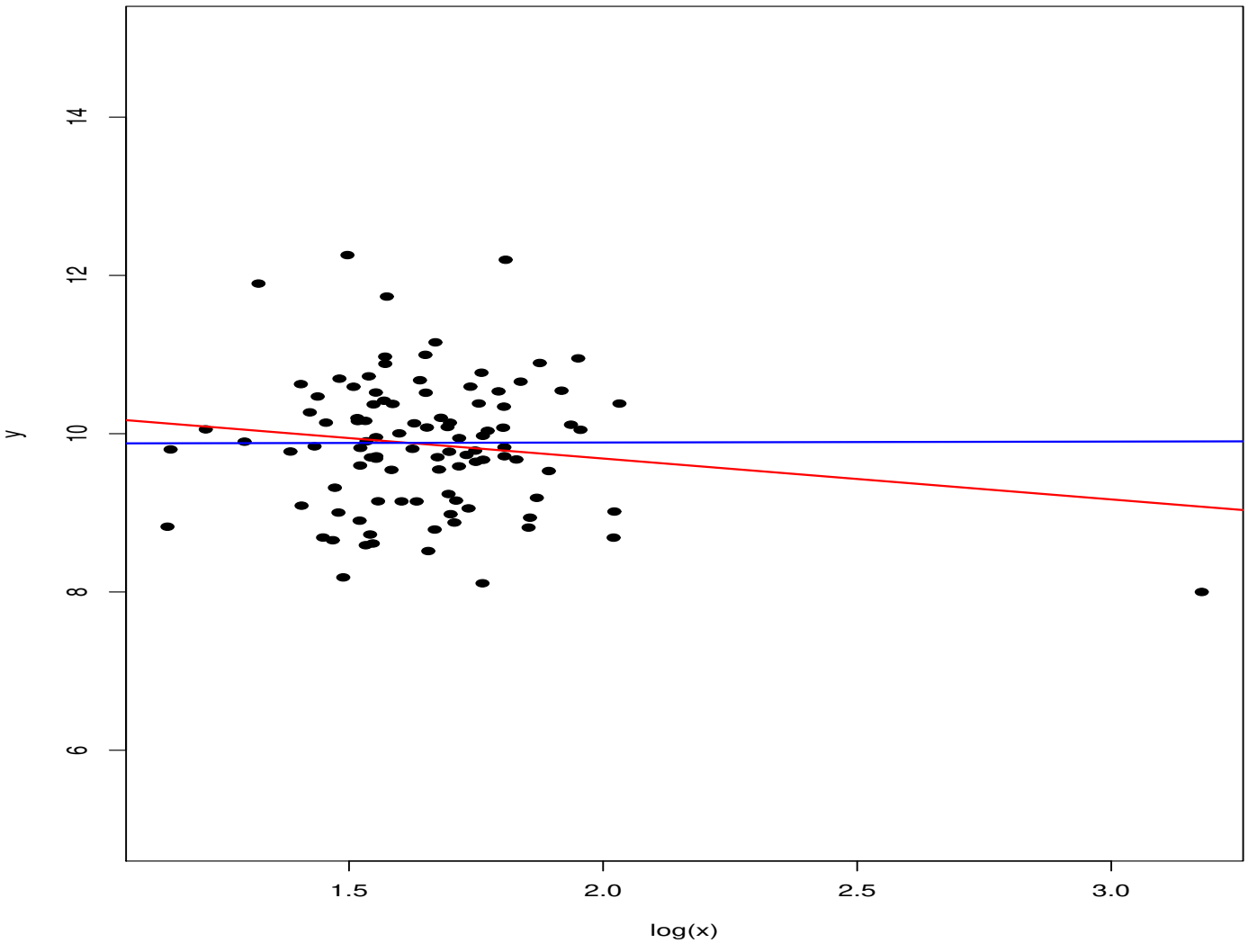
Vanessa McNealis

What relationship do we see if we omit the 1 outlying point?

```
Call:
lm(formula = y[x < 20] ~ x[x < 20])

Residuals:
     Min       1Q   Median       3Q      Max
-1.77342 -0.66950  0.01179  0.49940  2.37098

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.893887   0.475902   20.79   <2e-16 ***
x[x < 20]   -0.001794   0.089756   -0.02    0.984
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

Residual standard error: 0.8264 on 97 degrees of freedom
Multiple R-Squared: 4.119e-06, Adjusted R-squared: -0.01031
F-statistic: 0.0003995 on 1 and 97 DF,  p-value: 0.984
```

Model-checking: outliers

Vanessa McNealis



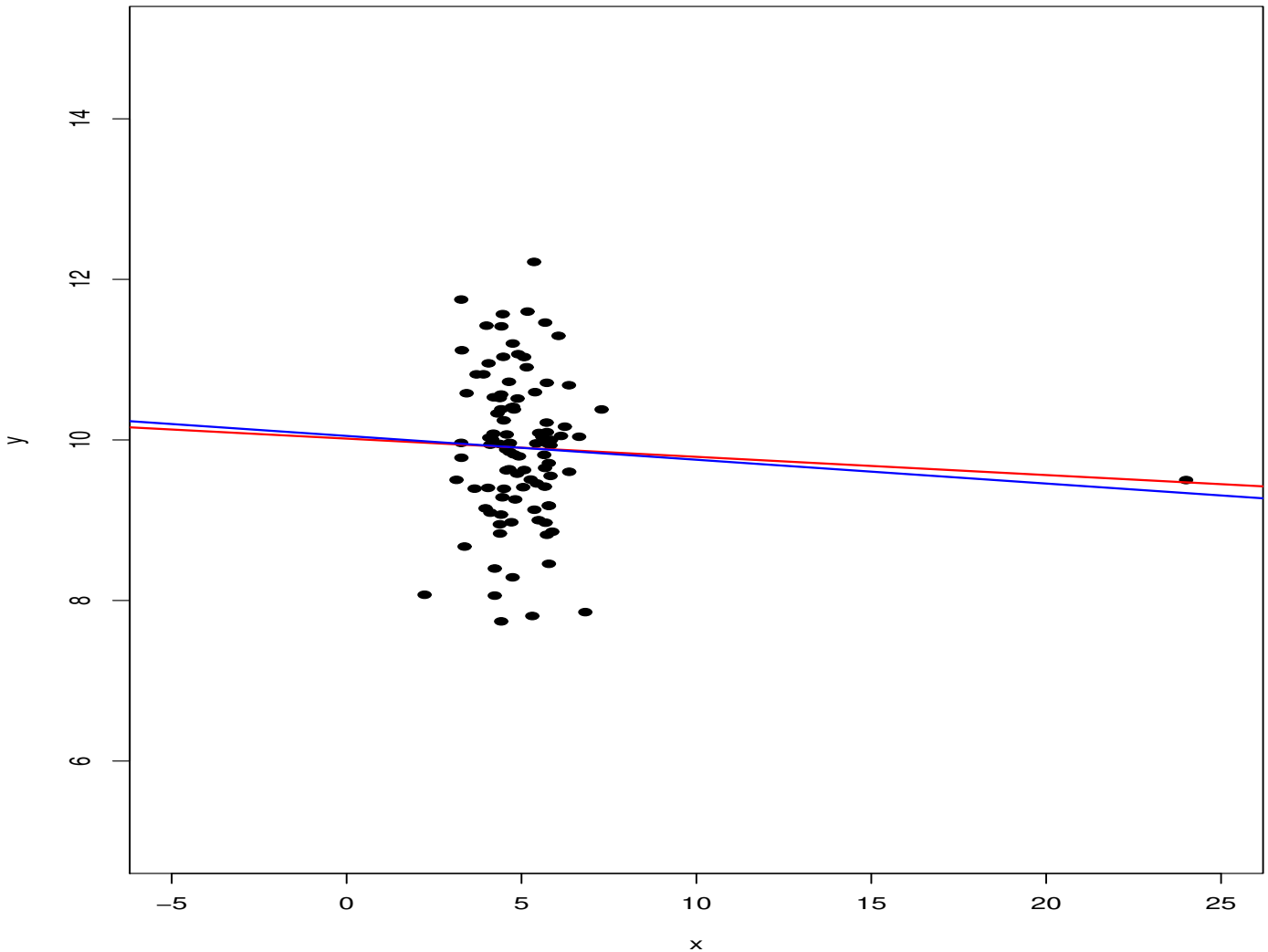
Model-checking: outliers

What relationship do we see if we log-transform x?

```
Call:
lm(formula = y ~ log.x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.76620 -0.64566 -0.03259  0.51960  2.41261

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.7194     0.5943  18.037   <2e-16 ***
log.x        -0.5167     0.3561  -1.451     0.15
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8348 on 98 degrees of freedom
Multiple R-Squared: 0.02103,    Adjusted R-squared: 0.01104
F-statistic: 2.105 on 1 and 98 DF,  p-value: 0.15
```

Model-checking: outliers

Vanessa McNealis



Model-checking: outliers

Not all data points that are *univariate* outliers influence our regression line.

```
> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.17617 -0.52683  0.03945  0.60676  2.32295

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.01616    0.23978  41.772   <2e-16 ***
x           -0.02269    0.04377  -0.518    0.605
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9187 on 98 degrees of freedom
Multiple R-Squared: 0.002735,   Adjusted R-squared: -0.007442
F-statistic: 0.2687 on 1 and 98 DF,  p-value: 0.6054
```

Model-checking: outliers

Possible explanations for outliers:

- Bad data may result from explainable events, e.g. malfunction of measuring instrument, incorrect recording of data. In this case, try to retrieve the correct value. If that isn't possible, it may be best to discard the observation.
- Inadequacies in the model. The model may fail to fit the data well for certain values of the predictor. In this case it could be disastrous to simply discard outliers.
- Sampling of observations in the tail of the distribution (deliberate or otherwise). This may be especially likely to happen if the outcome arises from a heavy-tailed distribution.

In short: if the outlier arises from a known error, try to fix it. Otherwise, keep it and consider transformation and/or a different model form.

Model-checking: outliers

Possible explanations for outliers:

- Bad data may result from explainable events, e.g. malfunction of measuring instrument, incorrect recording of data. In this case, try to retrieve the correct value. If that isn't possible, it may be best to discard the observation.
- Inadequacies in the model. The model may fail to fit the data well for certain values of the predictor. In this case it could be disastrous to simply discard outliers.
- Sampling of observations in the tail of the distribution (deliberate or otherwise). This may be especially likely to happen if the outcome arises from a heavy-tailed distribution.

In short: if the outlier arises from a known error, try to fix it. Otherwise, keep it and consider transformation and/or a different model form.

Model-checking: outliers

Possible explanations for outliers:

- Bad data may result from explainable events, e.g. malfunction of measuring instrument, incorrect recording of data. In this case, try to retrieve the correct value. If that isn't possible, it may be best to discard the observation.
- Inadequacies in the model. The model may fail to fit the data well for certain values of the predictor. In this case it could be disastrous to simply discard outliers.
- Sampling of observations in the tail of the distribution (deliberate or otherwise). This may be especially likely to happen if the outcome arises from a heavy-tailed distribution.

In short: if the outlier arises from a known error, try to fix it. Otherwise, keep it and consider transformation and/or a different model form.

Model-checking: outliers

Possible explanations for outliers:

- Bad data may result from explainable events, e.g. malfunction of measuring instrument, incorrect recording of data. In this case, try to retrieve the correct value. If that isn't possible, it may be best to discard the observation.
- Inadequacies in the model. The model may fail to fit the data well for certain values of the predictor. In this case it could be disastrous to simply discard outliers.
- Sampling of observations in the tail of the distribution (deliberate or otherwise). This may be especially likely to happen if the outcome arises from a heavy-tailed distribution.

In short: if the outlier arises from a known error, try to fix it. Otherwise, keep it and consider transformation and/or a different model form.

Analysis of Variance (ANOVA)

We are going to see how the variability of the observations $Y_1, \ldots, Y_n$ can be decomposed under the linear regression model. This decomposition is essential to perform hypotheses tests in multiple regression models. Furthermore, this decomposition appears in the output of any software, even when one fits a simple linear regression model.

ANOVA

First, let's note that $\bar{\hat{Y}} = \bar{Y}$. Then, the variability of the observations $Y_1, \ldots, Y_n$ can be decomposed as follows:

$$\begin{aligned}
\sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\
&= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \\
&\qquad 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \\
&= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.
\end{aligned}$$

If we let
$SS_{Total} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$, $SS_{Model} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ and
$SS_{Error} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, then, we have

$$SS_{Total} = SS_{Model} + SS_{Error}.$$

ANOVA

The equation on the previous slide is very important in regression.

- The **total sum of squares** SS_T (also called SS_{yy}) quantifies the variability of the Y_i's. The equation on the previous tells us that this variability can be decomposed into two components: SS_M and SS_E.
- The **model sum of squares** SS_M represents the variability of the $\hat{Y}_i$'s. Since the $\hat{Y}_i$'s vary only with the x_i's, SS_M represents the variability of the Y_i explained by the fact that all observations don't have the same values of x.
- The **error sum of square** SS_E measures the variability of the errors $Y_i - \hat{Y}_i$. This variability is explained by the fact that the value of Y is not totally explained by x (i.e. by the regression model).

ANOVA

In summary, we have:

$$\boxed{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,}$$

$$\boxed{SS_T = SS_M + SS_E,}$$

$$\boxed{\text{variability of } Y_i = \text{variability due to } x_i + \text{variability of errors.}}$$

ANOVA

Vanessa McNealis

Each sum of square can be associated with a degree of freedom:

SS_M:	1 degree of freedom (one explanatory variable)
SS_E:	$n-2$ degrees of freedom (n obs. minus 2 estimated params.)
SS_T:	$n-1$ degrees of freedom (n obs. minus an average)

ANOVA – Hypothesis test of $H_0 : \beta_1 = 0$

There are 3 approaches for testing this hypothesis:

Approach 1

This approach is simple, and consists simply in computing a confidence interval for β_1 at level $100\%(1-\alpha)$. If the value zero doesn't belong to the CI, we reject H_0 at level α. Note that this approach is valid only for two-sided alternative hypotheses.

Hypothesis test of $H_0 : \beta_1 = 0$

Approach 2

We can use the classic approach, based on the distribution of $\hat{\beta}_1$ under H_0. In this case, the test statistics is:

$$t = \frac{\hat{\beta}_1}{\hat{se}(\hat{\beta}_1)}.$$

For a two-sided test, a large value of $|t|$ leads to the rejection of H_0. More precisely, we reject H_0 at level α if $|t| \geq t_{\alpha/2;n-2}$. If α is not given, the p-value of the test is $\mathrm{P}(|t_{n-2}| \geq |t|) = 2\mathrm{P}(t_{n-2} \geq |t|)$.

Hypothesis test of $H_0 : \beta_1 = 0$

Approach 3

This approach is completely equivalent to approach 2 in the case of a simple linear regression model (but not in a multiple regression model!). The null hypothesis is

$$H_0 : Y_i = \beta_0 + \varepsilon_i, \quad \text{vs } H_1 : Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

The idea is to compare SS_M to SS_E to see if the explanatory variable explains a significant part of the total variability of the Y's. To do such a test, we use the following **ANOVA table**:

Source of variation	Sum of squares	df	Mean squares	F value
Model	SS_M	1	$MS_M = SS_M/1$	$F = MS_M/MS_E$
Error	SS_E	$n-2$	$MS_E = SS_E/(n-2)$	
Total	SS_T	$n-1$		

Hypothesis test of $H_0 : \beta_1 = 0$

Looking at the table, we should reject H_0 whenever SS_M is large relative to SS_E. This also corresponds to a large value of the statistic F. The test statistic is then:

$$F = MS_M / MS_E.$$

Under H_0, this statistic follows the distribution of Fisher–Snedecor $F_{1,n-2}$. The p-value of the test is $\mathrm{P}(F_{1,n-2} \geq F)$.

ANOVA for a factor variable

Any analysis-of-variance problem can be treated as a regression problem in which all the covariates are factor variables.
In a single factor analysis of variance model with k levels, we have:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1, \ldots, k, j = 1, \ldots, n_i$$

where

- μ is the grand mean
- τ_i is the effect associated to factor level i
- ϵ_{ij} is an error term with zero mean associated to individual j in group i

It is usually of interest to test

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_k$$

vs

$$H_1 : \tau_i \neq \tau_j \text{ for at least one pair } (i, j).$$

ANOVA for a factor variable

It turns out that the model can be reformulated as

$$Y_i = \beta_0 + \beta_1(x_i == 2) + \ldots + \beta_{k-1}(x_i == k) + \epsilon_i$$

where

- β_0 is the mean of level $x = 1$
- β_{l-1} is the effect associated with factor level $x = l$, $l = 2, \ldots, k$
- $(x_i == l)$ is a binary indicator taking value 1 if $x_i = l$, and 0 otherwise

Thus, testing for the absence of effect is equivalent to testing whether all the above regression coefficients are equal to 0.
Note: This model corresponds to a **multiple linear regression model**.

ANOVA for a factor variable

Approach 3 can be used when x is a factor variable with $k >= 2$ levels taking values 1,2,...,k. The null hypothesis is

$$H_0 : Y_i = \beta_0 + \varepsilon_i, \quad \text{vs}$$

$$H_1 : Y_i = \beta_0 + \beta_1(x_i == 2) + ... + \beta_{k-1}(x_i == k) + \varepsilon_i.$$

Again, we compare SS_M to SS_E to see if x explains a significant part of the total variability of the Y's. Now we use the following ANOVA table:

Source of variation	Sum of squares	df	Mean squares	F value
Model	SS_M	$k-1$	$MS_M = SS_M/(k-1)$	$F = MS_M/MS_E$
Error	SS_E	$n-k$	$MS_E = SS_E/(n-k)$	
Total	SS_T	$n-1$		

ANOVA for a factor variable

Looking at the table, we should reject H_0 whenever SS_M is large relative to SS_E. This also corresponds to a large value of the statistic F. The test statistic is then:

$$F = MS_M / MS_E.$$

Under H_0, this statistic follows the distribution of Fisher–Snedecor $F_{k-1,n-k}$. The p-value of the test is $\mathrm{P}(F_{k-1,n-k} \geq F)$.

Understanding the ANOVA F-statistic

Suppose that we have $k = 3$ treatment groups in a Completely Randomized Design, with sample sizes $n_1 = n_2 = n_3 = 6$. Suppose first that the treatment means are all equal to zero, that is

$$\mu_1 = \mu_2 = \mu_3 = 0$$

and that the treatment group variance parameter σ^2 is equal to 1. A typical data set is displayed below:

							$\bar{y}_i$	s_i^2
Tx 1	-0.88	0.24	-0.46	0.78	-0.47	-0.38	-0.195	0.358
Tx 2	-0.75	0.11	0.64	1.98	-1.03	1.84	0.465	1.611
Tx 3	1.38	1.20	0.42	0.05	-1.29	-0.04	0.287	0.939

yielding $\bar{y} = 0.186$, and

$$s_P^2 = \frac{1}{n-k}\sum_{i=1}^{k}(n_i - 1)s_i^2 = 0.969.$$

Understanding the ANOVA F-statistic

For these data, we use the definitions

$$SS_M = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2 = 1.399 \qquad SS_E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 = 14.539$$

and

$$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = 15.938$$

where $\bar{y}$ is the average over all observations and $\bar{y}_i$ is the average of observations under treatment i, $i = 1, \ldots, k$ so that the equation $SS_T = SS_M + SS_E$ holds. For the F-statistic, we have

$$F = \frac{MS_M}{MS_E} = \frac{SS_M/(k-1)}{SS_E/(n-k)} = \frac{1.399/2}{14.539/15} = 0.722$$

Understanding the ANOVA F-statistic

To complete the test, we compare this with the $1-\alpha$ probability point of the Fisher-F distribution with $(k-1, n-k) = (2, 15)$ degrees of freedom. With $\alpha = 0.05$, we see that

$$F_\alpha(2, 15) = 3.68$$

and we **do not reject** the ANOVA F-test null hypothesis

$$H_0 \quad : \quad \mu_1 = \mu_2 = \mu_3.$$

This is the **correct** conclusion, as in fact all the true treatment means are zero. Thus a **small** value of the test statistic F supports H_0.

Understanding the ANOVA F-statistic

Now suppose that, in fact,

$$\mu_1 = 0 \qquad \mu_2 = 10 \qquad \mu_3 = 20.$$

The equivalent data set to the one above but with the treatment means changed in this way takes the form

							$\bar{y}_i$	s_i^2
Tx 1	-0.88	0.24	-0.46	0.78	-0.47	-0.38	-0.195	0.358
Tx 2	9.25	10.11	10.64	11.98	8.97	11.84	10.465	1.611
Tx 3	21.38	21.20	20.42	20.05	18.71	19.96	20.287	0.939

yielding $\bar{y} = 10.186$, and

$$s_P^2 = \frac{1}{n-k}\sum_{i=1}^{k}(n_i - 1)s_i^2 = 0.969.$$

.

Understanding the ANOVA F-statistic

Note that the sample means have changed accordingly, but that the sample variances **have not changed at all**. On further calculation, we have

$$SS_M = 1259.199 \qquad SS_E = 14.539 \qquad SS_T = 1273.738$$

so that

$$MS_M = \frac{1259.199}{2} = 629.600 \qquad MS_E = \frac{14.539}{15} = 0.969$$

so that

$$F = \frac{629.600}{0.969} = 649.570.$$

We again compare this with $F_\alpha(2, 15) = 3.68$ (the critical value, C_R), and notice that the F statistic is **much larger** than this critical value. The test statistic thus lies within the rejection region, and hence we **reject $\mathbf{H_0}$**.

Understanding the ANOVA F-statistic

This example illustrates that SS_M measures the variability **between** means across the treatment groups, whereas SS_E measures the variability **within** treatment groups, allowing for the possibility that the treatment means may be different. The quantity SS_T measures the total amount of variability; in the first example $SS_T = SS_M + SS_E$ gives

$$15.938 = 1.399 + 14.539$$

so most of the variability is contributed by SS_E, whereas in the second example, we have

$$1273.738 = 1259.199 + 14.539$$

and most of the variability is contributed by SS_M.

Goodness-of-fit

To measure the goodness of fit of a model, we define the R^2 coefficient as

$$\begin{aligned} R^2 &= 1 - \frac{SS_E}{SS_T} \\ &= \frac{SS_M}{SS_T}. \end{aligned}$$

We can easily see that $0 \leq R^2 \leq 1$.

- When $R^2 = 0$, all the variability of Y is explained by the error, which means that the variable x doesn't explain at all the value of Y.
- When $R^2 = 1$, all the data points fall on the regression line, which means that the fit of the model is perfect.
- The R^2 coefficient can then be interpreted as the proportion of the variability of Y explained by the variable x.

The coefficient of correlation

To measure the *strength of linear association* between the two variables x and y we can use the

Pearson Product Moment Coefficient of Correlation

or *correlation coefficient*.

The coefficient, r, is defined by

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

where

$$SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad SS_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

The coefficient of correlation

To measure the *strength of linear association* between the two variables x and y we can use the

Pearson Product Moment Coefficient of Correlation

or *correlation coefficient*.

The coefficient, r, is defined by

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

where

$$SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad SS_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

The coefficient of correlation

Two equivalent formulae for the correlation coefficient are

$$r = \frac{\sum_{i=1}^{n}(x_i y_i) - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{SS_{xx} SS_{yy}}}$$

and

$$r = \widehat{\beta}_1 \sqrt{\frac{SS_{xx}}{SS_{yy}}}.$$

The coefficient of correlation

Note: $-1 \leq r \leq 1$.

- If r is close to 1, there is a strong linear relationship between x and y where y **increases** with x.
- If r is close to -1, there is a strong linear relationship between x and y where y **decreases** with x.

Note: In the model

$$y = \beta_0 + \beta_1 x$$

$\beta_1 = 0 \Longrightarrow r \approx 0$, so tests for $\beta_1 = 0$ can also be used to deduce a lack of correlation between the variables.

What can't we do?

We have learned how to relate two continuous variables using linear regression and correlation. When is this not enough?

- What if we observe a systematic trend (e.g., a U-shape) when we plot residuals against the covariate?
  - **Polynomial regression.**
- What if we observe that variance is not constant?
  - **Weighted least squares** allows us to reweight the data in such a way that inference is valid.
  - We can also sometimes use transformations to correct this.
- What if we want to explain the variation in response, y, using more than one covariate?
  - **Multiple linear regression.**

What can't we do?

We have learned how to relate two continuous variables using linear regression and correlation. When is this not enough?

- What if we observe a systematic trend (e.g., a U-shape) when we plot residuals against the covariate?
  - **Polynomial regression.**
- What if we observe that variance is not constant?
  - **Weighted least squares** allows us to reweight the data in such a way that inference is valid.
  - We can also sometimes use transformations to correct this.
- What if we want to explain the variation in response, y, using more than one covariate?
  - **Multiple linear regression.**

What can't we do?

We have learned how to relate two continuous variables using linear regression and correlation. When is this not enough?

- What if we observe a systematic trend (e.g., a U-shape) when we plot residuals against the covariate?
  - **Polynomial regression.**
- What if we observe that variance is not constant?
  - **Weighted least squares** allows us to reweight the data in such a way that inference is valid.
  - We can also sometimes use transformations to correct this.
- What if we want to explain the variation in response, y, using more than one covariate?
  - **Multiple linear regression.**

What can't we do?

We have learned how to relate two continuous variables using linear regression and correlation. When is this not enough?

- What if we observe a systematic trend (e.g., a U-shape) when we plot residuals against the covariate?
  - **Polynomial regression.**
- What if we observe that variance is not constant?
  - **Weighted least squares** allows us to reweight the data in such a way that inference is valid.
  - We can also sometimes use transformations to correct this.
- What if we want to explain the variation in response, y, using more than one covariate?
  - **Multiple linear regression.**

What can't we do?

- What if response is binary? Linear regression returns a predicted response that is continuous, and so this is inappropriate.
  - We can generalize the linear model to **logistic regression.**
- What if response is a Poisson count, or something else altogether?
  - Again, linear regression is inappropriate, but we can use generalizations of the linear model such as **log-linear regression** models.
- What if not all individuals are followed for the same length of time when examining a binary outcome (e.g., we are looking at time to remission, but participants drop out throughout the study so that time-to-remission is a censored variable)?
  - **Survival** models (e.g., Cox model).

What can't we do?

- What if response is binary? Linear regression returns a predicted response that is continuous, and so this is inappropriate.
  - We can generalize the linear model to **logistic regression.**
- What if response is a Poisson count, or something else altogether?
  - Again, linear regression is inappropriate, but we can use generalizations of the linear model such as **log-linear regression** models.
- What if not all individuals are followed for the same length of time when examining a binary outcome (e.g., we are looking at time to remission, but participants drop out throughout the study so that time-to-remission is a censored variable)?
  - **Survival** models (e.g., Cox model).

What can't we do?

- What if response is binary? Linear regression returns a predicted response that is continuous, and so this is inappropriate.
  - We can generalize the linear model to **logistic regression.**
- What if response is a Poisson count, or something else altogether?
  - Again, linear regression is inappropriate, but we can use generalizations of the linear model such as **log-linear regression** models.
- What if not all individuals are followed for the same length of time when examining a binary outcome (e.g., we are looking at time to remission, but participants drop out throughout the study so that time-to-remission is a censored variable)?
  - **Survival** models (e.g., Cox model).

What can't we do?

- What if response is binary? Linear regression returns a predicted response that is continuous, and so this is inappropriate.
  - We can generalize the linear model to **logistic regression.**
- What if response is a Poisson count, or something else altogether?
  - Again, linear regression is inappropriate, but we can use generalizations of the linear model such as **log-linear regression** models.
- What if not all individuals are followed for the same length of time when examining a binary outcome (e.g., we are looking at time to remission, but participants drop out throughout the study so that time-to-remission is a censored variable)?
  - **Survival** models (e.g., Cox model).

Supplementary Materials

Interpretation: log transformations of x

We may want to transform our data so that the influence of outlying values is reduced.

Let's consider a log (base b) transformation of the covariate: letting $x^{new} = \log_b(x)$. What effect does this have on our interpretation of the parameters?

Interpretation: log transformations of x

Recall the properties of logs:

For example, say we have $x^{new} = \log_2(x)$. Then the reverse transformation is $x = 2^{x^{new}}$.

This tells us that increasing x^{new} by 1 unit is equivalent to *doubling* the covariate x on the original scale, since for $x = 2^{x^{new}}$,

$$2^{(x^{new}+1)} = 2 \times 2^{x^{new}} = 2x.$$

So we would now **interpret β_1 as the expected change in response Y associated with a doubling of the covariate x.**

Interpretation: log transformations of x

Recall the properties of logs:

For example, say we have $x^{new} = \log_2(x)$. Then the reverse transformation is $x = 2^{x^{new}}$.

This tells us that increasing x^{new} by 1 unit is equivalent to *doubling* the covariate x on the original scale, since for $x = 2^{x^{new}}$,

$$2^{(x^{new}+1)} = 2 \times 2^{x^{new}} = 2x.$$

So we would now **interpret β_1 as the expected change in response Y associated with a doubling of the covariate x.**

Interpretation: log transformations of y

We may want to transform our data so that the distribution of Y (and therefore of the errors) is more symmetric (more Normal).

Let's consider a log (base b) transformation of the response: letting $y^{new} = \log_b(y)$. What effect does this have on our interpretation of the parameters?

Interpretation: log transformations of y

Again, recall the properties of logs:

For example, say we have $y^{new} = \log_{10}(y)$. Then the reverse transformation is $y = 10^{y^{new}}$.

This tells us that increasing x by 1 unit is equivalent to a change in y^{new} of β_1 units, on average. But what does that mean in terms of the change in response, y, on the original scale?

$$10^{(y^{new}+\beta_1)} = 10^{y^{new}} \times 10^{\beta_1} = y \times 10^{\beta_1}.$$

So we would now interpret the slope coefficient in the model as 10^{β_1} **is the factor by which response is expected to change the median of Y that is associated with a one-unit increase in the covariate x.**

Interpretation: log transformations of y

Again, recall the properties of logs:

For example, say we have $y^{new} = \log_{10}(y)$. Then the reverse transformation is $y = 10^{y^{new}}$.

This tells us that increasing x by 1 unit is equivalent to a change in y^{new} of β_1 units, on average. But what does that mean in terms of the change in response, y, on the original scale?

$$10^{(y^{new}+\beta_1)} = 10^{y^{new}} \times 10^{\beta_1} = y \times 10^{\beta_1}.$$

So we would now interpret the slope coefficient in the model as 10^{β_1} **is the factor by which response is expected to change the median of Y that is associated with a one-unit increase in the covariate x.**

Interpretation: summary

- Transforming x by taking a log (base b) means that we now interpret the slope parameter as the expected change in Y associated with a b-fold increase in x.
- Transforming y by taking a log (base b) means that we now interpret the slope parameter as a *median multiplicative* change of b^{β_1} in Y associated with a one unit increase in x.

ANOVA: example

A standard model of memory is that the degree to which the subject remembers verbal material is a function of the degree to which it was processed when it was initially presented. [Reference: Craik and Lockhart (1972). Levels of Processing: a framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11, 671-684.]

Experiment: Fifty subjects aged 55-65 were randomly assigned to one of five groups which carried out different memory tasks. The five groups included

- **Counting** group: asked to read through a list of words and simply count the number of letters in each word.
- **Rhyming** group: asked to read each word and think of a word that rhymed with it.
- **Adjective** group: had to give an adjective that could reasonably be used to modify each word on the list.
- **Imagery** group: instructed to try to form vivid images of each word.
- **Intentional** group: told to read through the list and to memorize the words for later recall.

ANOVA: example

After subjects had gone through the list of 27 items three times, they were given a sheet of paper and asked to write down all the words they could remember. The response data were the number of words recalled by each individual in each group, and are presented below:

Counting	Rhyming	Adjective	Imagery	Intentional
9	7	11	12	10
8	9	13	11	19
6	6	8	16	14
8	6	6	11	5
10	6	14	9	10
4	11	11	23	11
6	6	13	12	14
5	3	13	10	15
7	8	10	19	11
7	7	11	11	11

ANOVA: example

Research question: Does the level of processing required when material is processed affect how much material is remembered ?

Test a hypothesis to answer this question using an ANOVA F-test. Specifically form the ANOVA table, and report the result of the ANOVA F-test.

ANOVA: example

```
> mem.dat <- read.table("MemoryTask.txt",sep=" ",header=T)
> mem.dat$Memory.task <-  as.factor(mem.dat$Memory.task)
> summary(lm(mem.dat$Memory~mem.dat$Memory.task))

Call: lm(formula = mem.dat$Memory ~ mem.dat$Memory.task)

Residuals:
   Min     1Q Median     3Q    Max
 -7.00  -1.85  -0.45   2.00   9.60

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            7.0000     0.9835   7.117 6.83e-09 ***
mem.dat$Memory.task2  -0.1000     1.3909  -0.072 0.943004
mem.dat$Memory.task3   4.0000     1.3909   2.876 0.006138 **
mem.dat$Memory.task4   6.4000     1.3909   4.601 3.43e-05 ***
mem.dat$Memory.task5   5.0000     1.3909   3.595 0.000802 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.11 on 45 degrees of freedom
Multiple R-squared: 0.4468,     Adjusted R-squared: 0.3976
F-statistic: 9.085 on 4 and 45 DF,  p-value: 1.815e-05
```

ANOVA: example


```
…(from previous slide)…
Residual standard error: 3.11 on 45 degrees of freedom
Multiple R-squared: 0.4468,     Adjusted R-squared: 0.3976
F-statistic: 9.085 on 4 and 45 DF,  p-value: 1.815e-05

> anova(lm(mem.dat$Memory~mem.dat$Memory.task))
Analysis of Variance Table

Response: mem.dat$Memory
                    Df Sum Sq Mean Sq F value    Pr(>F)
mem.dat$Memory.task  4 351.52  87.880  9.0848 1.815e-05 ***
Residuals           45 435.30   9.673
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple linear regression

Now, several covariates $x_1, x_2, \ldots, x_p$ are used to explain the variation in the response y. We have the probabilistic model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p + \epsilon,$$

where ϵ is a random error term with zero mean and variance σ^2 (same as in the simple case). Terminology:

- β_0 - **Intercept** parameter
- $\beta_1, \ldots, \beta_p$ - **Coefficients** of $x_1, \ldots, x_p$, respectively

Interpreting regression coefficients

$$E[Y|x] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p$$

- β_0 is the expected value of Y in a group of individuals with $x_1 = 0, x_2 = 0, \ldots x_p = 0$.
- For $k = 1, \ldots, p$, β_k is the expected difference in response Y between two groups of individuals, where one group has covariate value $x_k^* = x_k + 1$ and the other group has covariate value x_k, while all other covariates are equal across groups.

Multiple linear regression: Example

The following data show the fuel consumption over the 50 United States and the District of Columbia. The goal of this example is to understand the relationship between `FuelC` and other variables that were recorded. [Reference: Weisberg (2014). *Applied linear regression* (4th edition), p. 15-17.]

- The number of licensed drivers in the state (`Drivers`)
- Per person personal income, in dollars (`Income`)
- Miles of Federal-aid highway miles in the state (`Miles`)
- Population aged 16 or over (`Pop`)
- Gasoline state tax rate, cents per gallon (`Tax`)

	Drivers	FuelC	Income	Miles	MPC	Pop	Tax
AL	3559897	2382507	23471	94440	12737.00	3451586	18.00
AK	472211	235400	30064	13628	7639.16	457728	8.00
AZ	3550367	2428430	25578	55245	9411.55	3907526	18.00
AR	1961883	1358174	22257	98132	11268.40	2072622	21.70
CA	21623793	14691753	32275	168771	8923.89	25599275	18.00
CO	3287922	2048664	32949	85854	9722.73	3322455	22.00

Multiple linear regression: Example

We compute additional variables from the variables in the data set:

- `Fuel`: 1000× `FuelC`/`Pop`
- `Dlic`: 1000× `Drivers`/`Pop`
- log(`Miles`): Natural logarithm of `Miles`

We will examine the model

$$\text{Fuel}_i = \beta_0 + \beta_1 \text{Tax}_i + \beta_2 \text{Dlic}_i + \beta_3 \text{Income}_i + \beta_4 \log(\text{Miles})_i + \epsilon_i.$$

Multiple linear regression: Example

Vanessa McNealis

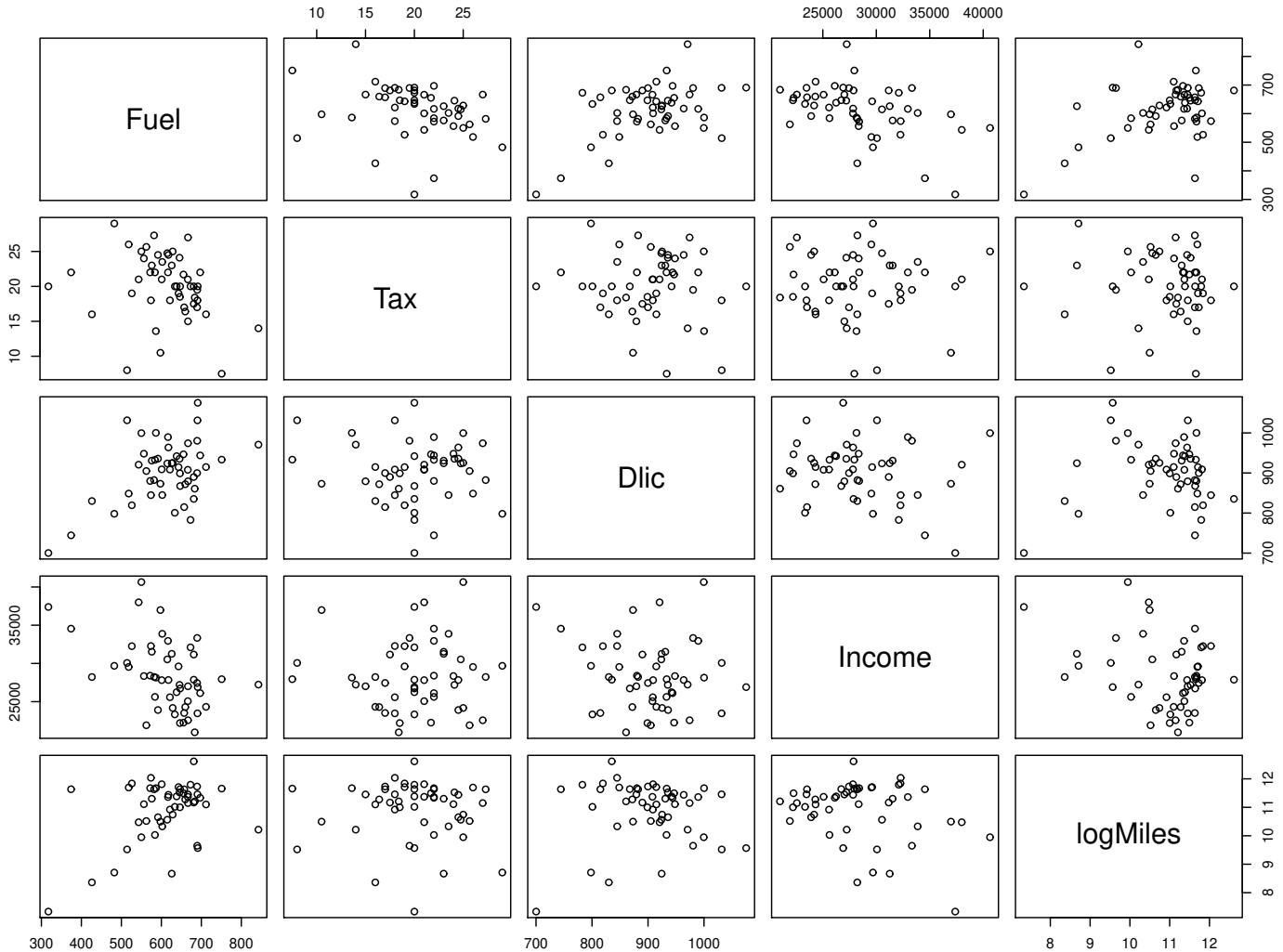


Figure: Scatterplot matrix for the fuel data

Multiple linear regression: Example

The marginal relationships between `Fuel` and each of the predictors are not sufficient to understand the *joint* effect of the predictors on the response. Instead, we can find the least-squares estimates of $\beta_0, \beta_1, \ldots, \beta_4$ in the model on slide 114:

```
> fit <- lm(Fuel ~ Tax + Dlic + Income + logMiles, data = fuel2001)
> summary(fit)

Call:
lm(formula = Fuel ~ Tax + Dlic + Income + logMiles, data = fuel2001)

Residuals:
     Min       1Q   Median       3Q      Max
-163.145  -33.039    5.895   31.989  183.499

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 154.192845 194.906161   0.791 0.432938
Tax          -4.227983   2.030121  -2.083 0.042873 *
Dlic          0.471871   0.128513   3.672 0.000626 ***
Income       -0.006135   0.002194  -2.797 0.007508 **
logMiles     26.755176   9.337374   2.865 0.006259 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 64.89 on 46 degrees of freedom      A
Multiple R-squared:  0.5105,    Adjusted R-squared:  0.4679  B
F-statistic: 11.99 on 4 and 46 DF,  p-value: 9.331e-07       C
```

Multiple linear regression: Example

From the regression output, we observe that:

- (A) The residual standard error is estimated to be

$$\hat{\sigma} = \sqrt{\frac{SS_E}{n-p}} = \sqrt{\frac{SS_E}{51-5}} = 64.89$$

- (B) The goodness-of-fit as measured by the multiple R^2 statistic is given by

$$R^2 = \frac{SS_M}{SS_T} = 1 - \frac{SS_E}{SS_T} = 0.5105.$$

- (C) The F statistic corresponding to the test $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ is given by

$$F = \frac{SS_M/4}{SS_E/(51-5)} = 11.99.$$

The null hypothesis is rejected at the level 5%, which leads us to conclude that the model is significant or in other words, the intercept-only model is not a reasonable simplification of the full model.

Multiple linear regression: Example

All the individual slopes are significantly different 0 at the 5% level. The T-statistics displayed in the fourth column of the summary are calculated as

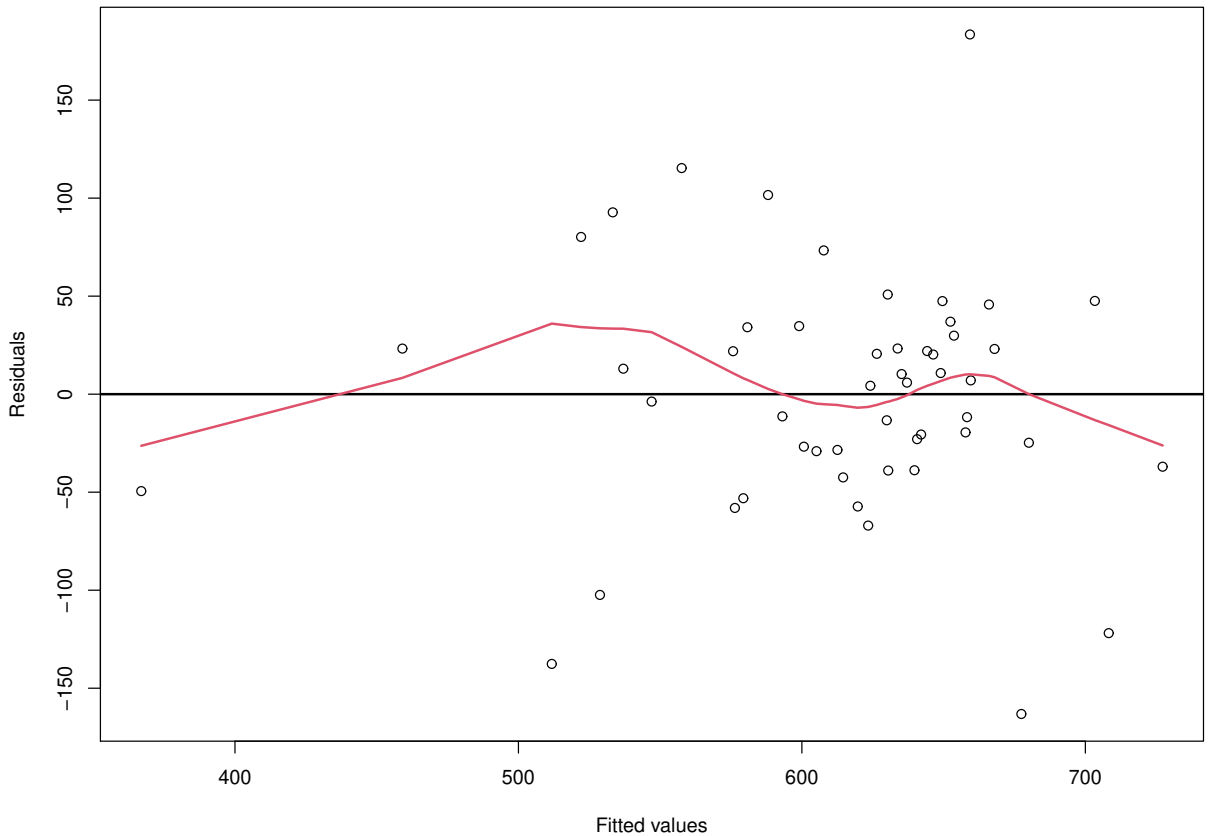
$$T_i = \frac{\hat{\beta}_i}{SE_{\hat{\beta}_i}} \overset{H_0}{\sim} t_{51-5},$$

where $SE_{\hat{\beta}_i}$ is displayed in column 3.
Interpretation:

- For every one cent-per-gallon increase in gasoline state tax rate, on average, the gasoline sold decreases by 4.22 gallons per person per in the year 2001, all other things being equal.

Multiple linear regression: Example

Vanessa McNealis



The plot of residuals against fitted values suggests a relatively good fit.