

# STAT 401A - Statistical Methods for Research Workers

## Simple linear regression

Jarad Niemi (Dr. J)

Iowa State University

last updated: October 14, 2014

# Simple Linear Regression

Recall the one-way ANOVA model:

$$Y_{ij} \stackrel{\text{ind}}{\sim} N(\mu_j, \sigma^2)$$

where  $Y_{ij}$  is the observation for individual  $i$  in group  $j$ .

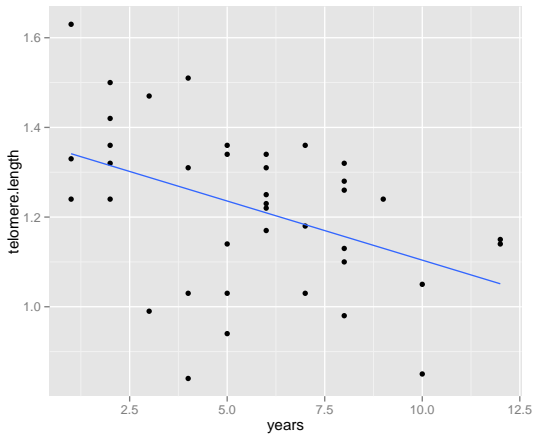
The **simple linear regression** model is

$$Y_i \stackrel{\text{ind}}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

where  $Y_i$  and  $X_i$  are the response and explanatory variable, respectively, for individual  $i$ .

Terminology (all of these are equivalent):

response	explanatory
outcome	covariate
dependent	independent
endogenous	exogenous



# Telomere length

<http://www.pnas.org/content/101/49/17312>

*People who are stressed over long periods tend to look haggard, and it is commonly thought that psychological stress leads to premature aging and the earlier onset of diseases of aging.*

...

*This design allowed us to examine the importance of perceived stress and measures of objective stress (caregiving status and chronicity of caregiving stress based on the number of years since a child's diagnosis).*

...

*Telomere length values were measured from DNA by a quantitative PCR assay that determines the relative ratio of telomere repeat copy number to single-copy gene copy number (T/S ratio) in experimental samples as compared with a reference DNA sample.*

# Interpretation

$$E[Y_i|X_i = x] = \beta_0 + \beta_1 x \quad V[Y_i|X_i = x] = \sigma^2$$

- If  $X_i = 0$ , then  $E[Y_i|X_i = 0] = \beta_0$ .

$\beta_0$  is the expected response when the explanatory variable is zero.

- If  $X_i$  increases from  $x$  to  $x + 1$ , then

$$\begin{array}{rcl} E[Y_i|X_i = x + 1] & = & \beta_0 + \beta_1 x + \beta_1 \\ - E[Y_i|X_i = x] & = & \beta_0 + \beta_1 x \\ \hline & = & \beta_1 \end{array}$$

$\beta_1$  is the expected increase in the response for each unit increase in the explanatory variable.

- $\sigma$  is the standard deviation of the response for a fixed value of the explanatory variable.

Remove the mean:

$$Y_i = \beta_0 + \beta_1 X_i + e_i \quad e_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

So the error is

$$e_i = Y_i - (\beta_0 + \beta_1 X_i)$$

which we approximate by the **residual**

$$r_i = \hat{e}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

The least squares, maximum likelihood, and Bayesian estimators are

$$\begin{aligned} \hat{\beta}_1 &= SXY / SXX \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\sigma}^2 &= SSE / (n - 2) \quad df = n - 2 \end{aligned}$$

$$\begin{aligned} \bar{X} &= \frac{1}{n} \sum_{i=1}^n X_i \\ \bar{Y} &= \frac{1}{n} \sum_{i=1}^n Y_i \end{aligned}$$

$$\begin{aligned} SXY &= \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \\ SXX &= \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X}) = \sum_{i=1}^n (X_i - \bar{X})^2 \\ SSE &= \sum_{i=1}^n r_i^2 \end{aligned}$$

How certain are we about  $\hat{\beta}_0$  and  $\hat{\beta}_1$  being equal to  $\beta_0$  and  $\beta_1$ ?

We quantify this uncertainty using their standard errors:

$$SE(\beta_0) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_X^2}} \quad df = n - 2$$

$$SE(\beta_1) = \hat{\sigma} \sqrt{\frac{1}{(n-1)s_X^2}} \quad df = n - 2$$

$$s_X^2 = SXX / (n - 1)$$

$$s_Y^2 = SYY / (n - 1)$$

$$SYY = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$r_{XY} = \frac{SXY / (n-1)}{s_X s_Y}$$

$$R^2 = r_{XY}^2$$

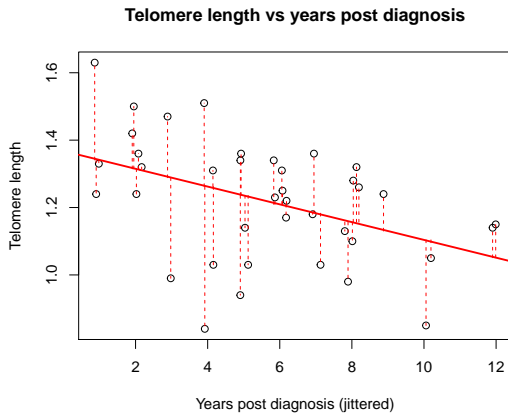
correlation coefficient

$$= \frac{SST - SSE}{SST}$$

coefficient of determination

$$SST = SYY = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

The coefficient of determination ( $R^2$ ) is the proportion of the total response variation explained by the explanatory variable(s).





# Pvalues and confidence interval

We can compute two-sided pvalues via

$$2P\left(t_{n-2} < -\left|\frac{\hat{\beta}_0}{SE(\beta_0)}\right|\right) \quad \text{and} \quad 2P\left(t_{n-2} < -\left|\frac{\hat{\beta}_1}{SE(\beta_1)}\right|\right)$$

These test the null hypothesis that the corresponding parameter is zero.

We can construct  $100(1 - \alpha)\%$  two-sided confidence intervals via

$$\hat{\beta}_0 \pm t_{n-2}(1 - \alpha/2)SE(\beta_0) \quad \text{and} \quad \hat{\beta}_1 \pm t_{n-2}(1 - \alpha/2)SE(\beta_1)$$

These provide ranges of the parameters consistent with the data.

```
DATA t;
  INFILE 'telomeres.csv' DSD FIRSTOBS=2;
  INPUT years length;
```

```
PROC CORR DATA=t;
  VAR length;
  WITH years;
RUN;
```

#### The CORR Procedure

```
1 With Variables:   years
1   Variables:     length
```

#### Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
years	39	5.58974	2.93543	218.00000	1.00000	12.00000
length	39	1.22026	0.17977	47.59000	0.84000	1.63000

Pearson Correlation Coefficients, N = 39  
 Prob > |r| under H0: Rho=0

```
length
years   -0.43065
        0.0062
```

```
PROC GLM DATA=t;
  MODEL length = years / SOLUTION CLPARM;
  RUN;
```

The GLM Procedure

Number of Observations Read	39
Number of Observations Used	39

Dependent Variable: length

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.22776588	0.22776588	8.42	0.0062
Error	37	1.00033156	0.02703599		
Corrected Total	38	1.22809744			

R-Square	Coeff Var	Root MSE	length Mean
0.185462	13.47473	0.164426	1.220256

Source	DF	Type I SS	Mean Square	F Value	Pr > F
years	1	0.22776588	0.22776588	8.42	0.0062

Source	DF	Type III SS	Mean Square	F Value	Pr > F
years	1	0.22776588	0.22776588	8.42	0.0062

Parameter	Estimate	Standard Error	t Value	Pr >  t	95% Confidence Limits
Intercept	1.367682067	0.05721112	23.91	<.0001	1.251761335 1.483602799
years	-0.026374315	0.00908674	-2.90	0.0062	-0.044785794 -0.007962836

# Regression in R

```
m = lm(telomere.length~years, Telomeres)
with(Telomeres, cor(telomere.length,years))
```

```
[1] -0.4307
```

```
anova(m)
```

Analysis of Variance Table

Response: telomere.length

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
years	1	0.228	0.228	8.42	0.0062 **
Residuals	37	1.000	0.027		

---

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

# Regression in R

```
m = lm(telomere.length~years, Telomeres)
summary(m)
```

Call:

```
lm(formula = telomere.length ~ years, data = Telomeres)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.4222	-0.0854	0.0206	0.1074	0.2887

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.36768	0.05721	23.9	<2e-16 ***
years	-0.02637	0.00909	-2.9	0.0062 **

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

Residual standard error: 0.164 on 37 degrees of freedom

Multiple R-squared: 0.185, Adjusted R-squared: 0.163

F-statistic: 8.42 on 1 and 37 DF, p-value: 0.0062

```
confint(m)
```

	2.5 %	97.5 %
(Intercept)	1.25176	1.483603
years	-0.04479	-0.007963

# Conclusion

Telomere length at the time of diagnosis of a child's chronic illness is estimated to be 1.37 with a 95% confidence interval of (1.25, 1.48). For each year increase since diagnosis, the length decreases by 0.026 with a 95% confidence interval of (0.008, 0.045). The proportional of variability in telomere length described by years since diagnosis is 18.5%.

<http://www.pnas.org/content/101/49/17312>

*The zero-order correlation between chronicity of caregiving [years] and mean telomere length,  $r$ , is  $-0.445$  ( $P < 0.01$ ). [ $R^2 = 0.198$  was shown in the plot.]*

**Remark** I'm guessing our analysis and that reported in the paper don't match exactly due to a discrepancy in the data.

# Summary

- The **simple linear regression** model is

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

where  $Y_i$  and  $X_i$  are the response and explanatory variable, respectively, for individual  $i$ .

- Know how to use SAS/R to obtain  $\hat{\beta}_0$ ,  $\hat{\beta}_1$ ,  $\hat{\sigma}^2$ ,  $R^2$ , pvalues, CIs, etc.
- Interpret SAS output
  - At a value of zero for the explanatory variable ( $X_i = 0$ ),  $\beta_0$  is the expected value for the response ( $Y_i$ ).
  - For each unit increase in the explanatory variable value,  $\beta_1$  is the expected increase in the response.
  - At a constant value of the explanatory variable,  $\sigma^2$  is the variance of the responses.
  - The coefficient of determination ( $R^2$ ) is the percentage of the total response variation explained by the explanatory variable(s).

## What is $E[Y|X = x]$ ?

We know  $\beta_0 = E[Y|X = 0]$ , but what about  $X = x$ ?

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

which we can estimate via

$$E[\widehat{Y|X = x}] = \hat{\beta}_0 + \hat{\beta}_1 x$$

but there is uncertainty in both  $\beta_0$  and  $\beta_1$ . So the standard error of  $E[Y|X = x]$  is

$$SE(E[Y|X = x]) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(\bar{X} - x)^2}{(n-1)s_X^2}}$$

and a  $100(1 - \alpha)\%$  confidence interval is

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{n-2}(1 - \alpha/2)SE(E[Y|X = x])$$



## What do we predict about $Y$ at $X = x$ ?

On the last slide, we calculated  $E[Y|X = x]$  and it's uncertainty, but if we are trying to predict a new observation, we need to account for the sampling variability  $\sigma^2$ . Thus a prediction about  $Y$  at a new  $X = x$  is still

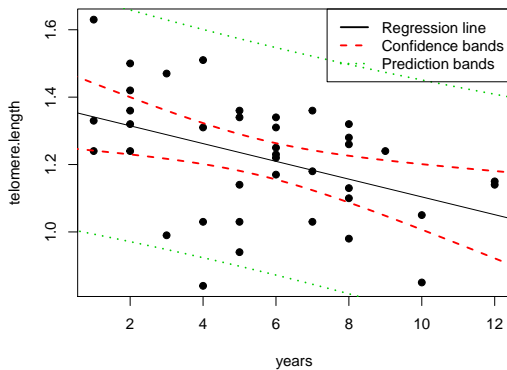
$$\text{Pred}(Y|X = x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

but the uncertainty includes the variability due to  $\sigma^2$ . So the standard error of  $\text{Pred}(Y|X = x)$  is

$$SE(\text{Pred}(Y|X = x)) = \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(\bar{X} - x)^2}{(n-1)s_X^2}}$$

and a  $100(1 - \alpha)\%$  confidence interval is

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{n-2}(1 - \alpha/2) SE(\text{Pred}(Y|X = x)).$$



# Testing Composite hypotheses

Comparing two models

- $H_0$  : (reduced)
- $H_1$  : (full)

Do the following

1. Calculate extra sum of squares.
2. Calculate extra degrees of freedom
3. Calculate

$$\text{F-statistic} = \frac{\text{Extra sum of squares} / \text{Extra degrees of freedom}}{\hat{\sigma}_{full}^2}$$

4. Compare this to an F-distribution with

- numerator degrees of freedom = extra degrees of freedom
- denominator degrees of freedom = degrees of freedom in estimating  $\hat{\sigma}_{full}^2$

# Simple Linear Regression

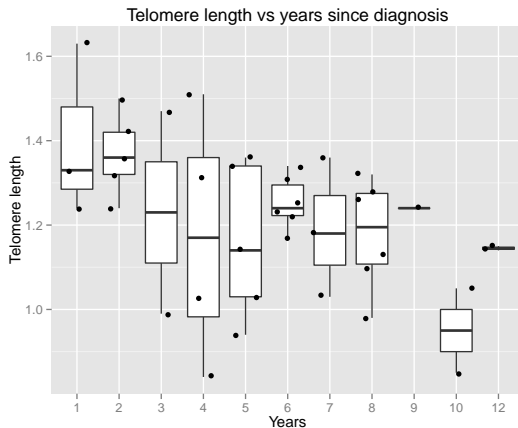
Two models:

ANOVA:  $Y_{ij} \overset{\text{ind}}{\sim} N(\mu_i, \sigma^2)$  (full)

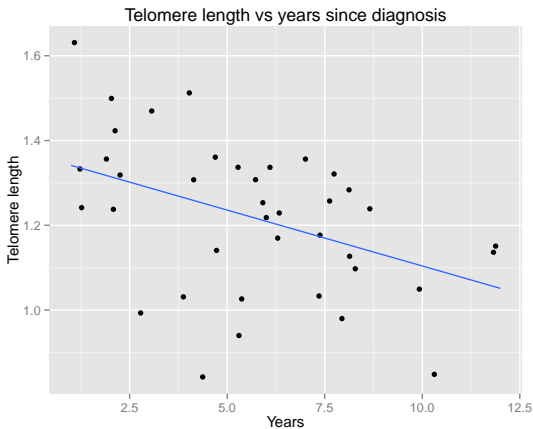
Regression:  $Y_{ij} \overset{\text{ind}}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$  (reduced)

- Regression model is reduced:
  - ANOVA has  $J$  parameters for the mean
  - Regression has 2 parameters for the mean
  - $\mu_i = \beta_0 + \beta_1 X_i$
- Small pvalues indicate a lack-of-fit, i.e. the reduced model is not adequate.
- Lack-of-fit F-test requires multiple observations at a few  $X_i$  values.

# Telomere length



# Telomere length



# SAS code

```
DATA t;
  INFILE 'telomeres.csv' DSD FIRSTOBS=2;
  INPUT years length;

PROC REG DATA=t;
  MODEL length = years / CLB LACKFIT;
  RUN;
```

The REG Procedure  
 Model: MODEL1  
 Dependent Variable: length

Number of Observations Read	39
Number of Observations Used	39

## Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.22777	0.22777	8.42	0.0062
Error	37	1.00033	0.02704		
Lack of Fit	9	0.18223	0.02025	0.69	0.7093
Pure Error	28	0.81810	0.02922		
Corrected Total	38	1.22810			

Indicates no evidence for a lack of fit, i.e. regression seems adequate.

```
m_anova = lm(telomere.length ~ as.factor(years), Telomeres)
m_reg   = lm(telomere.length ~ years, Telomeres)
anova(m_reg, m_anova)
```

Analysis of Variance Table

Model 1: telomere.length ~ years

Model 2: telomere.length ~ as.factor(years)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	37	1.000				
2	28	0.818	9	0.182	0.69	0.71

No evidence of a lack of fit.



# Summary

- Lack-of-fit F-test tests the assumption of linearity
- Needs multiple observations at various explanatory variable values
- Small pvalue indicates a lack-of-fit, i.e. means are not linear
  - Transform response, e.g. log
  - Transform explanatory variable
  - Add other explanatory variables

# Regression

The simpler linear regression model is

$$Y_i \stackrel{\text{ind}}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

this can be rewritten as

$$Y_i = \beta_0 + \beta_1 X_i + e_i \quad e_i \stackrel{\text{ind}}{\sim} N(0, \sigma^2)$$

where we estimate the errors via the residuals

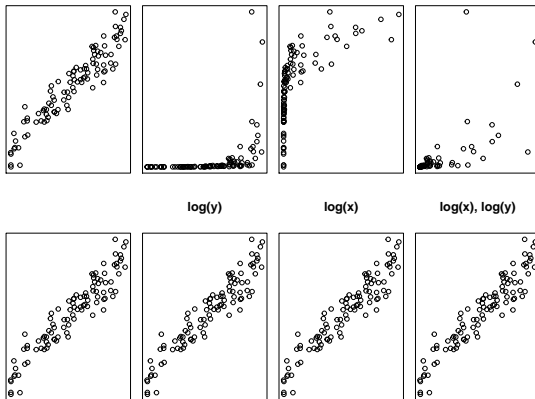
$$r_i = \hat{e}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i).$$

Key assumptions are:

- Linearity between mean response and explanatory variable
- Normality of the errors
- Constant variance of the errors
- Independence between observations

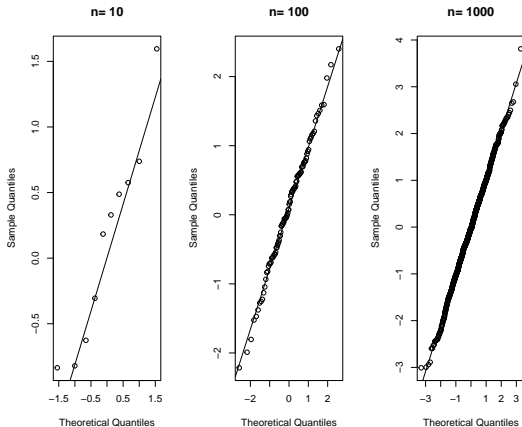
# Linearity

Assess using scatterplots of transformed response vs transformed explanatory variable:



# Normality

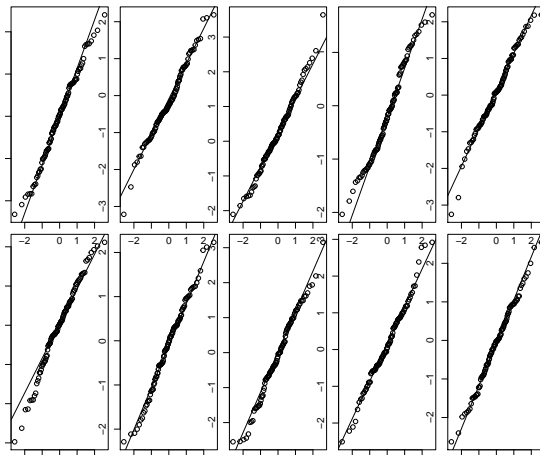
These are normal.



SAS swaps the x and y axes

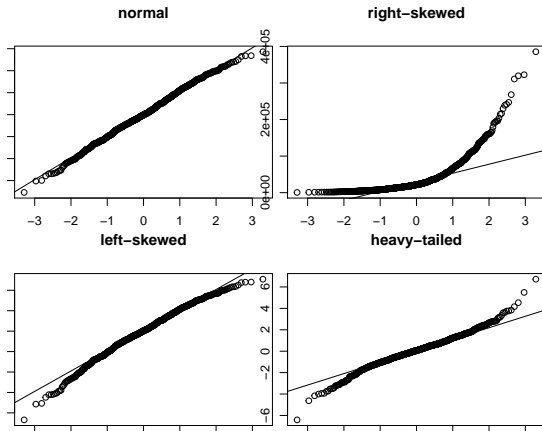
# Normality

These are normal.



SAS swaps the x and y axes

# Normality

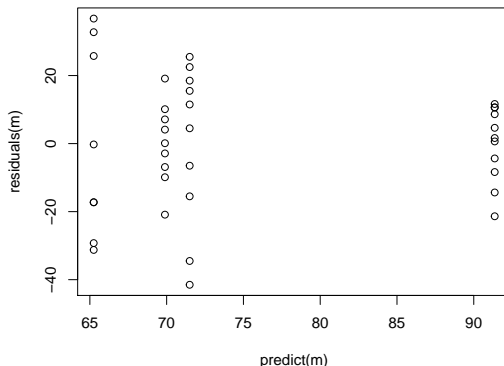


SAS swaps the x and y axes

# Constant variance

Most common non-constant variance is when the variance increases with the mean

Red Dye 40 residuals vs fitted values



# Independence

Lack of independence includes

- Cluster effect
- Serial correlation
- Spatial association

Make plots of residuals vs relevant explanatory variables and look for patterns, e.g.

- Residuals vs groups (prefer blocking)
- Residuals vs time (or observation number)
- Residuals vs spatial variable



# Summary

Often the best strategy is graphical exploration of the data, here are some relevant graphs:

- transformed response vs transformed explanatory
- transformed response vs transformed explanatory
- qqplot of residuals
- residual vs fitted value
- residual vs explanatory
- residual vs observation number
- residual vs any other variable