

BIM 105

Probability and Statistics for Biomedical Engineers

David M. Rocke

Department of Biomedical Engineering

Example Problem

In an accelerated life test, units are operated under extreme conditions until failure.

In one such test, 12 motors were operated under high temperature conditions.

The ambient temperatures (in $^{\circ}\text{C}$) and lifetimes (in hours) are presented in the following table:

Temperature Lifetime

40 851

45 635

50 764

55 708

60 469

65 661

70 586

75 371

80 337

85 245

90 129

95 158

100 hours ~ 4 days

1000 hours ~ 40 days

$40\text{ }^{\circ}\text{C} = 104\text{ }^{\circ}\text{F}$

$95\text{ }^{\circ}\text{C} = 203\text{ }^{\circ}\text{F}$

$120\text{ }^{\circ}\text{C} = 248\text{ }^{\circ}\text{F}$

- Construct a scatterplot of lifetime (y) versus temperature (x). Verify that a linear model is appropriate.
- Compute the least-squares line for predicting lifetime from temperature.
- Compute the fitted value and the residual for each point.
- If the temperature is increased by 5°C , by how much would you predict the lifetime to increase or decrease?
- Predict the lifetime for a temperature of 73°C .
- Should the least-squares line be used to predict the lifetime for a temperature of 120°C ? If so, predict the lifetime. If not, explain why not.
- For what temperature would you predict a lifetime of 500 hours?

```
>> scatter(Temperature,Lifetime,'filled')
>> xlabel('Temperature')
>> ylabel('Lifetime')
>> title('Accelerated Life Test')
>> lsline
```

```
>> mod1 = fitlm(Temperature,Lifetime)
```

Linear regression model:

$y \sim 1 + x_1$

Estimated Coefficients:

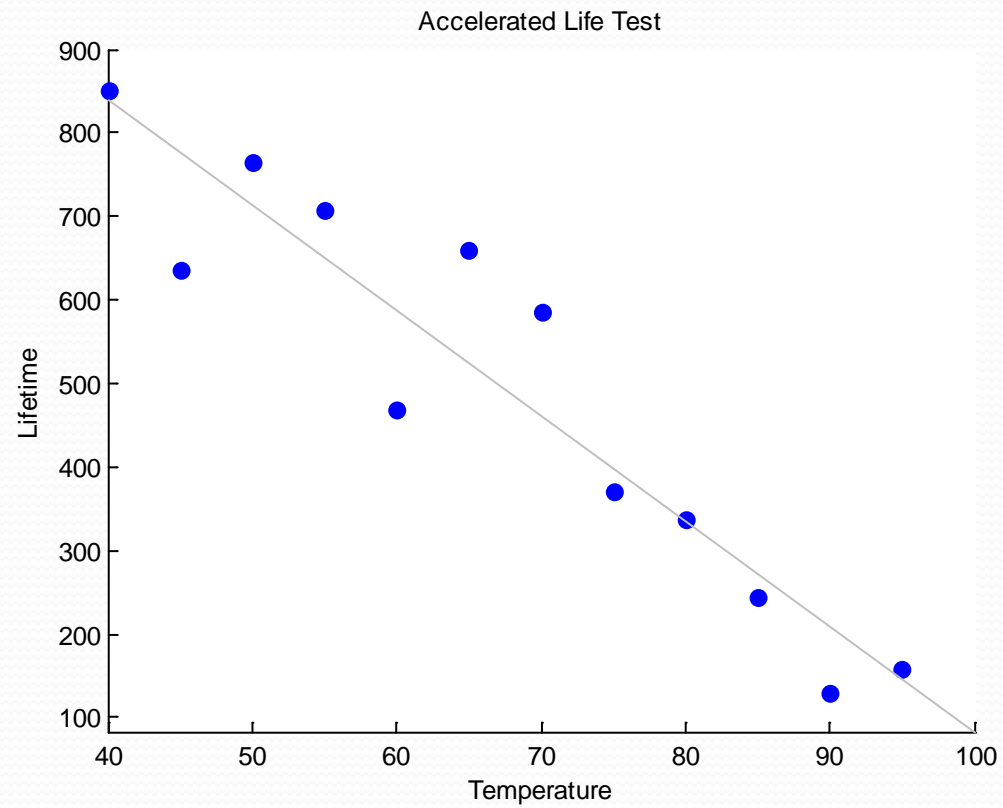
	Estimate	SE	tStat	pValue
(Intercept)	1344.1	105.72	12.714	1.6936e-07
x1	-12.611	1.5174	-8.3112	8.4166e-06

Number of observations: 12, Error degrees of freedom: 10

Root Mean Squared Error: 90.7

R-squared: 0.874, Adjusted R-Squared 0.861

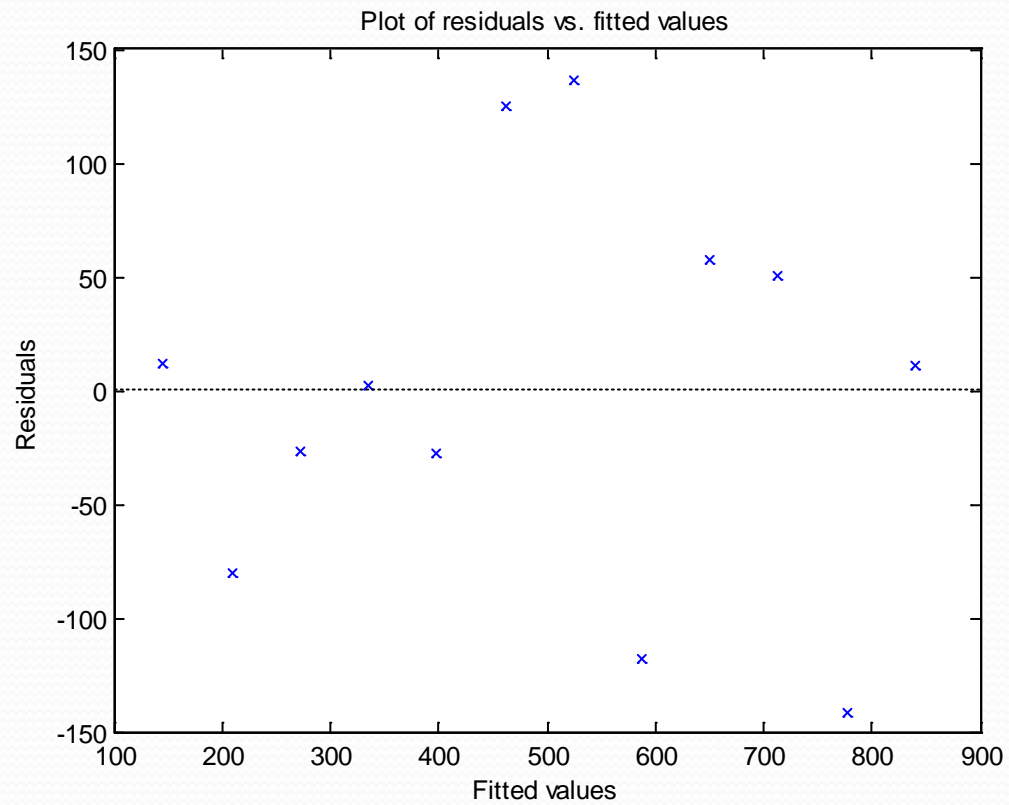
F-statistic vs. constant model: 69.1, p-value = 8.42e-06



```
>> [Temperature Lifetime mod1.Fitted mod1.Residuals.Raw]
```

40.0000	851.0000	839.6410	11.3590
45.0000	635.0000	776.5851	-141.5851
50.0000	764.0000	713.5291	50.4709
55.0000	708.0000	650.4732	57.5268
60.0000	469.0000	587.4172	-118.4172
65.0000	661.0000	524.3613	136.6387
70.0000	586.0000	461.3054	124.6946
75.0000	371.0000	398.2494	-27.2494
80.0000	337.0000	335.1935	1.8065
85.0000	245.0000	272.1375	-27.1375
90.0000	129.0000	209.0816	-80.0816
95.0000	158.0000	146.0256	11.9744

```
>> plotResiduals(mod1,'fitted')
```



	Estimate	SE	tStat	pValue
(Intercept)	1344.1	105.72	12.714	1.6936e-07
x1	-12.611	1.5174	-8.3112	8.4166e-06

If the temperature is increased by 5°C, by how much would you predict the lifetime to increase or decrease?

Decrease by $12.611 \times 5 = 63.1$ hours

Predict the lifetime for a temperature of 73°C.

$$1344.1 - (12.611)(73) = 423.5$$

Should the least-squares line be used to predict the lifetime for a temperature of 120°C? If so, predict the lifetime. If not, explain why not.

No. Too far out of range. Might not be linear. Anyway the prediction is negative (-169 hours), which is impossible.

For what temperature would you predict a lifetime of 500 hours?

$$(500 - 1344.1)/(-12.611) = 66.93 \text{ } ^\circ\text{C}$$

Quantitative Prediction

- Regression analysis is the statistical name for the prediction of one quantitative variable (fasting blood glucose level) from another (body mass index)
- Items of interest include whether there is in fact a relationship and what the expected change is in one variable when the other changes.
- A linear model is a prediction equation that is linear in the parameters. The simplest example is $y = a + bx$

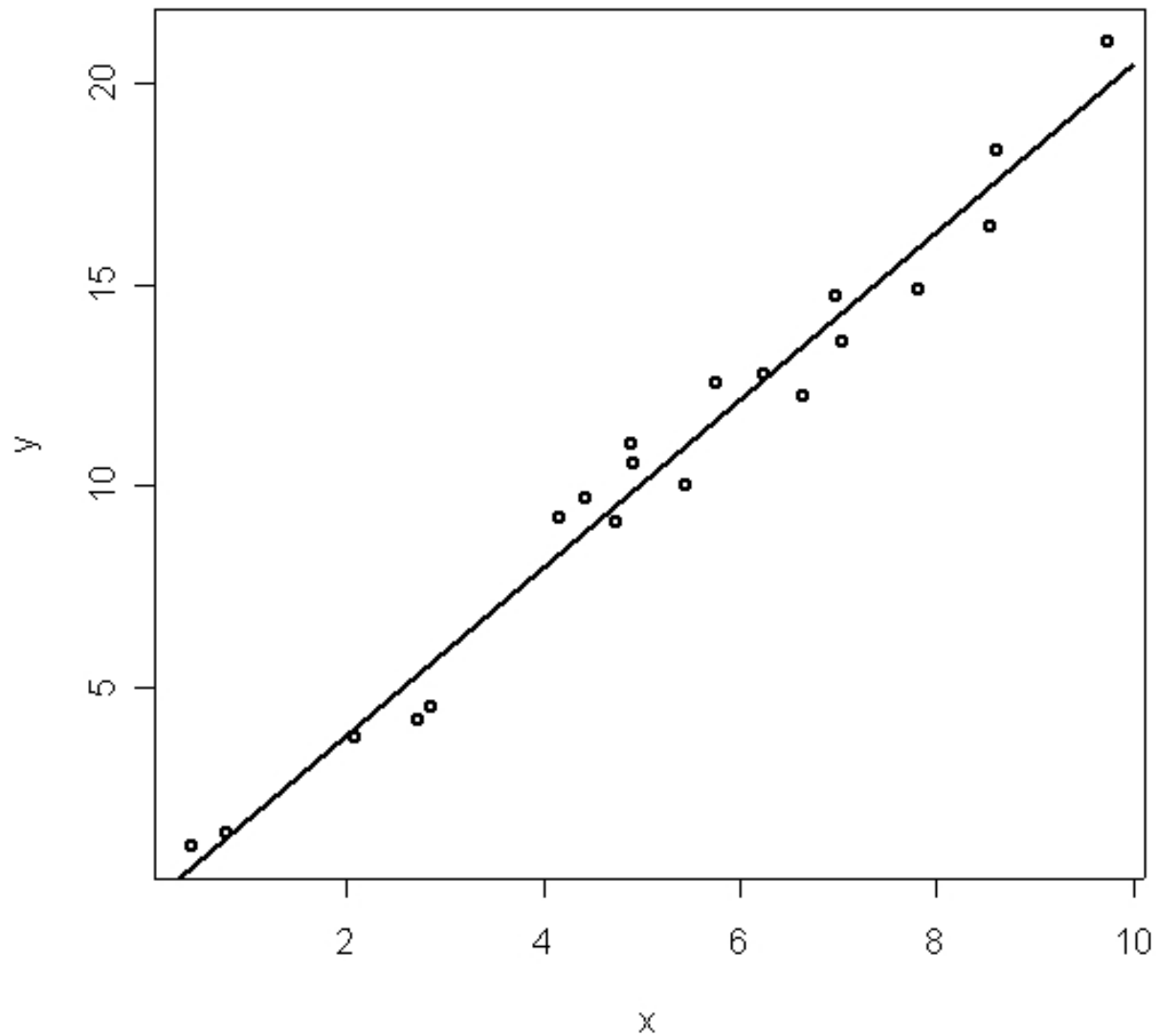
Assumptions

- Inference about whether there is a real relationship or not is dependent on a number of assumptions, many of which can be checked
- When these assumptions are substantially incorrect, alterations in method can sometimes rescue the analysis
- No assumption is ever exactly correct

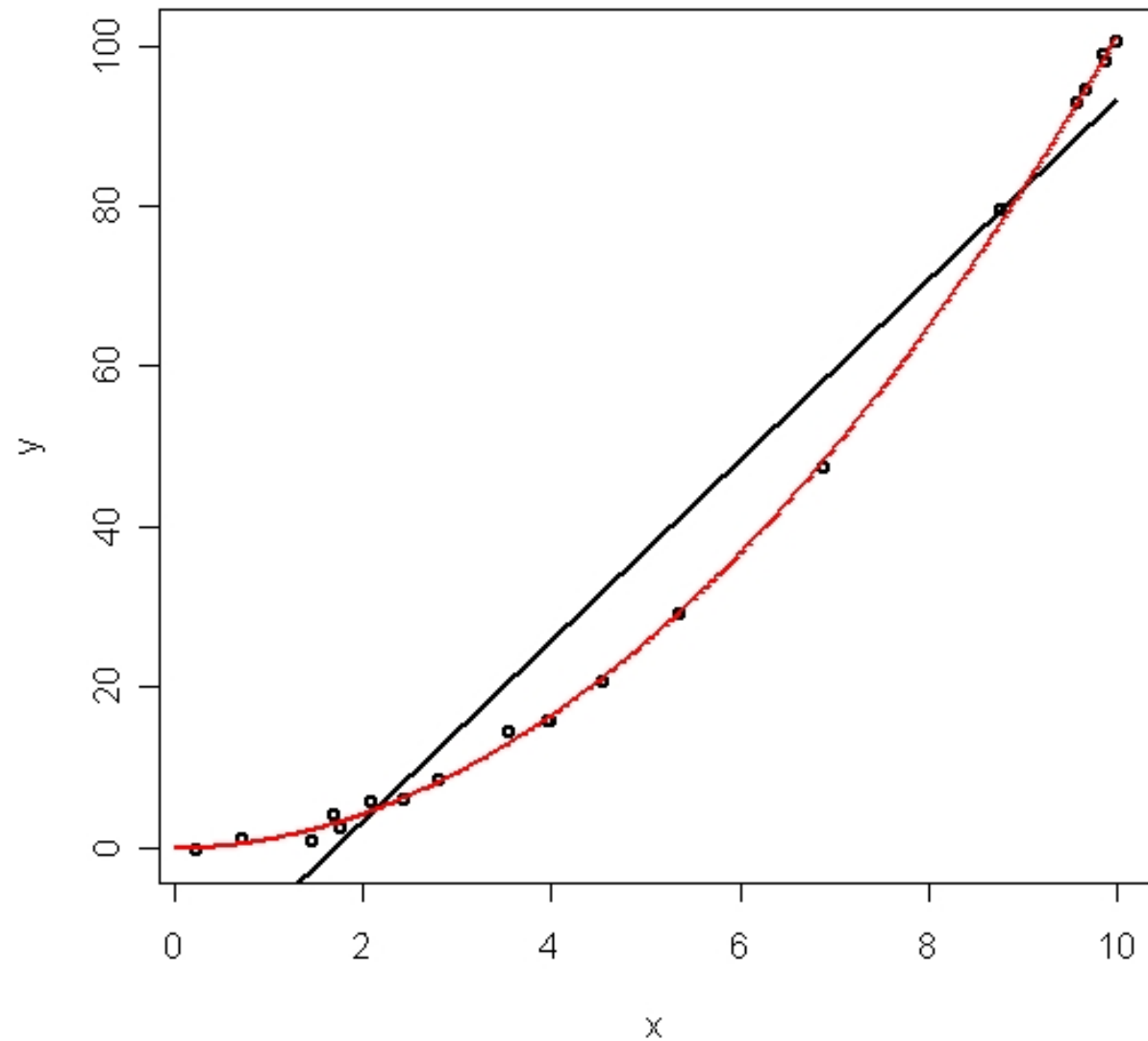
Linearity


- This is the most important assumption
- If x is the predictor, and y is the response, then we assume that the average response for a given value of x is a linear function of x
- $E(y) = a + bx$
- $y = a + bx + \varepsilon$
- ε is the *error* or variability

Regression when the Assumptions are Satisfied



Regression with nonlinearity

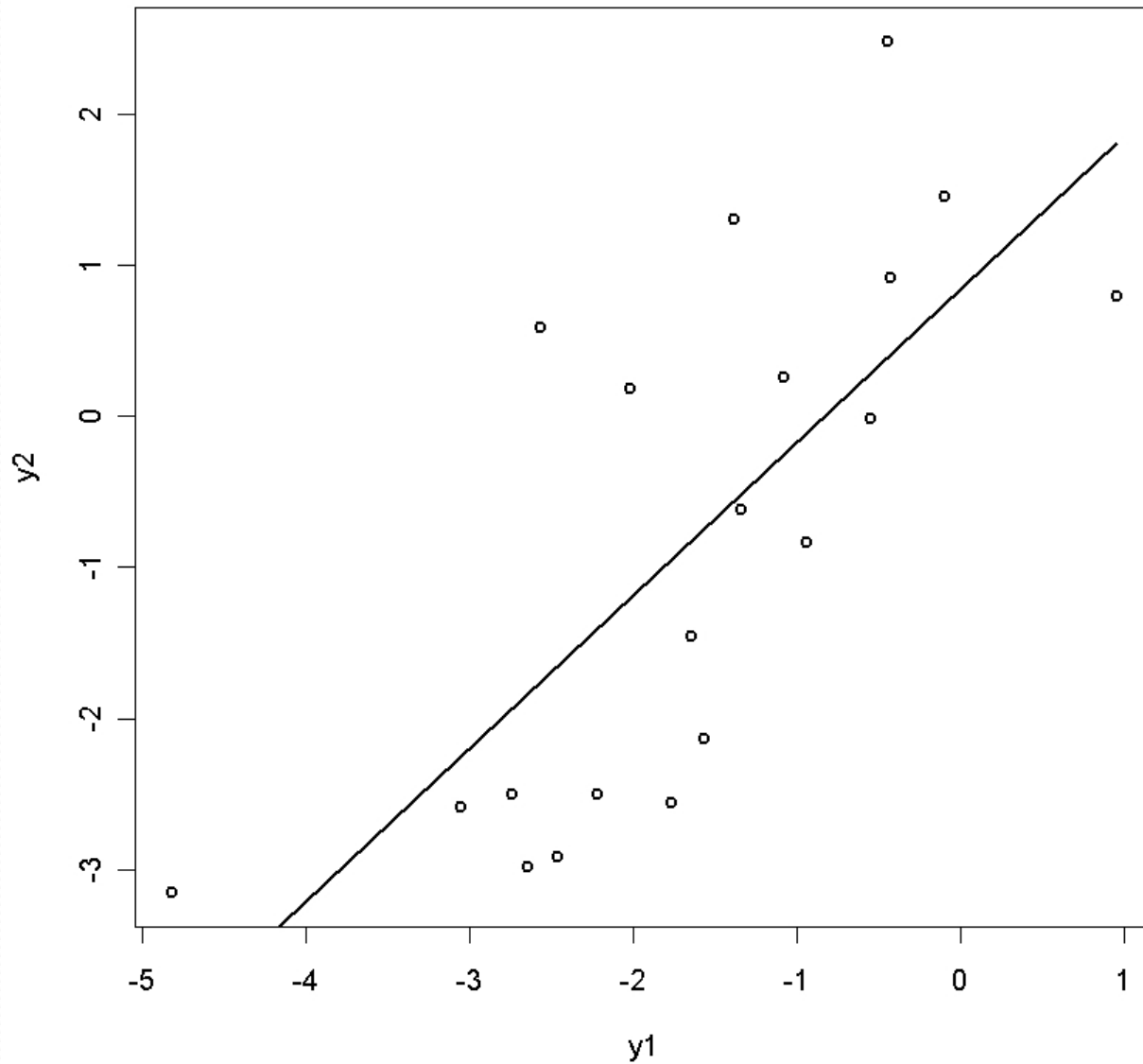


- 
- In general, it is important to get the model right, and the most important of these issues is that the mean function looks like it is specified
 - If a linear function does not fit, various types of curves can be used, but what is used should fit the data
 - Otherwise predictions are biased

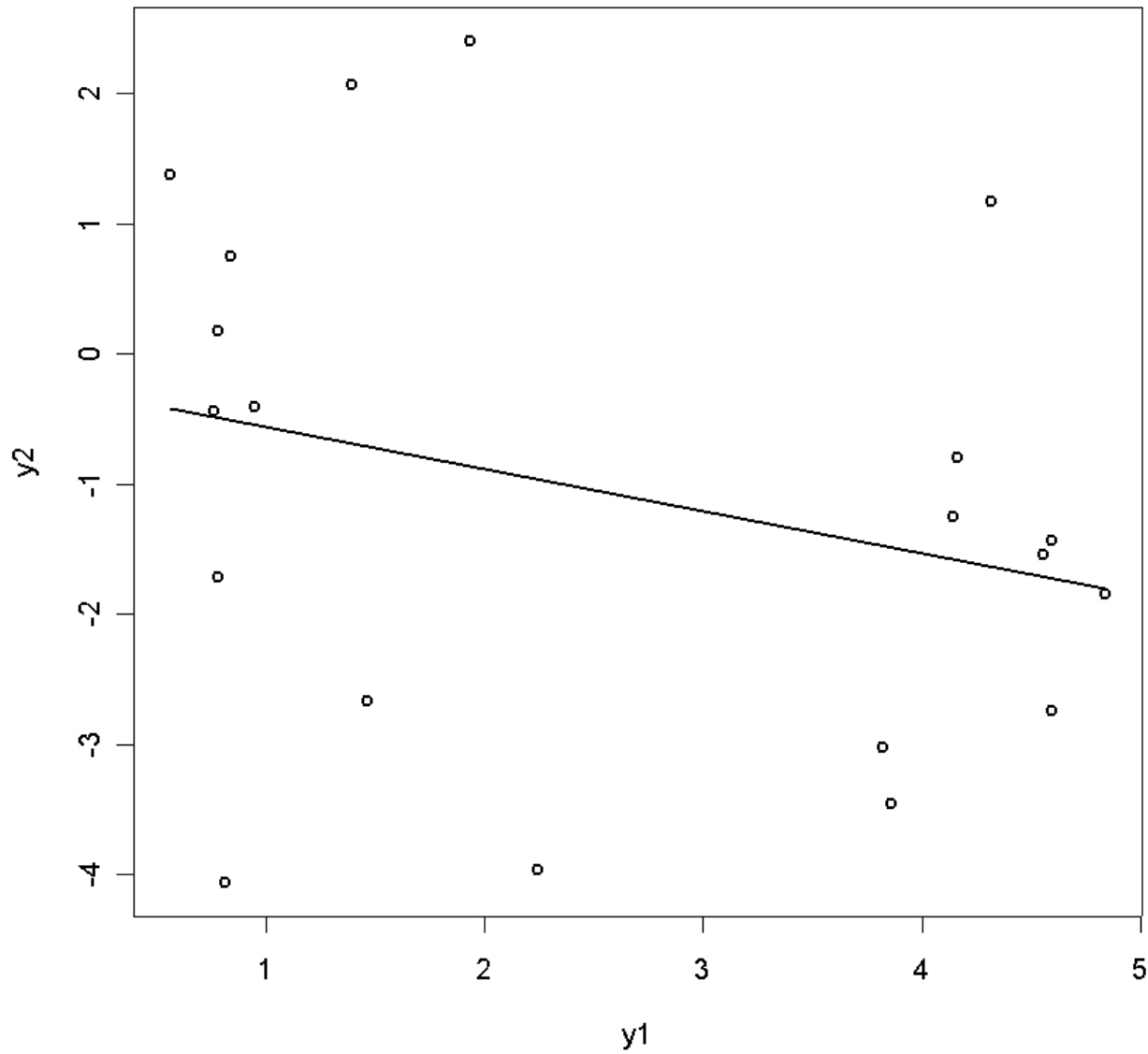
Independence

- It is assumed that different observations are statistically independent
- If this is not the case inference and prediction can be completely wrong
- There may appear to be a relationship even though there is not
- Randomization and then controlling the treatment assignment prevents this in general

Lack of Independence



Lack of Independence



- Note no true relationship between x and y
- These data were generated as follows:

$$x_1 = y_1 = 0$$

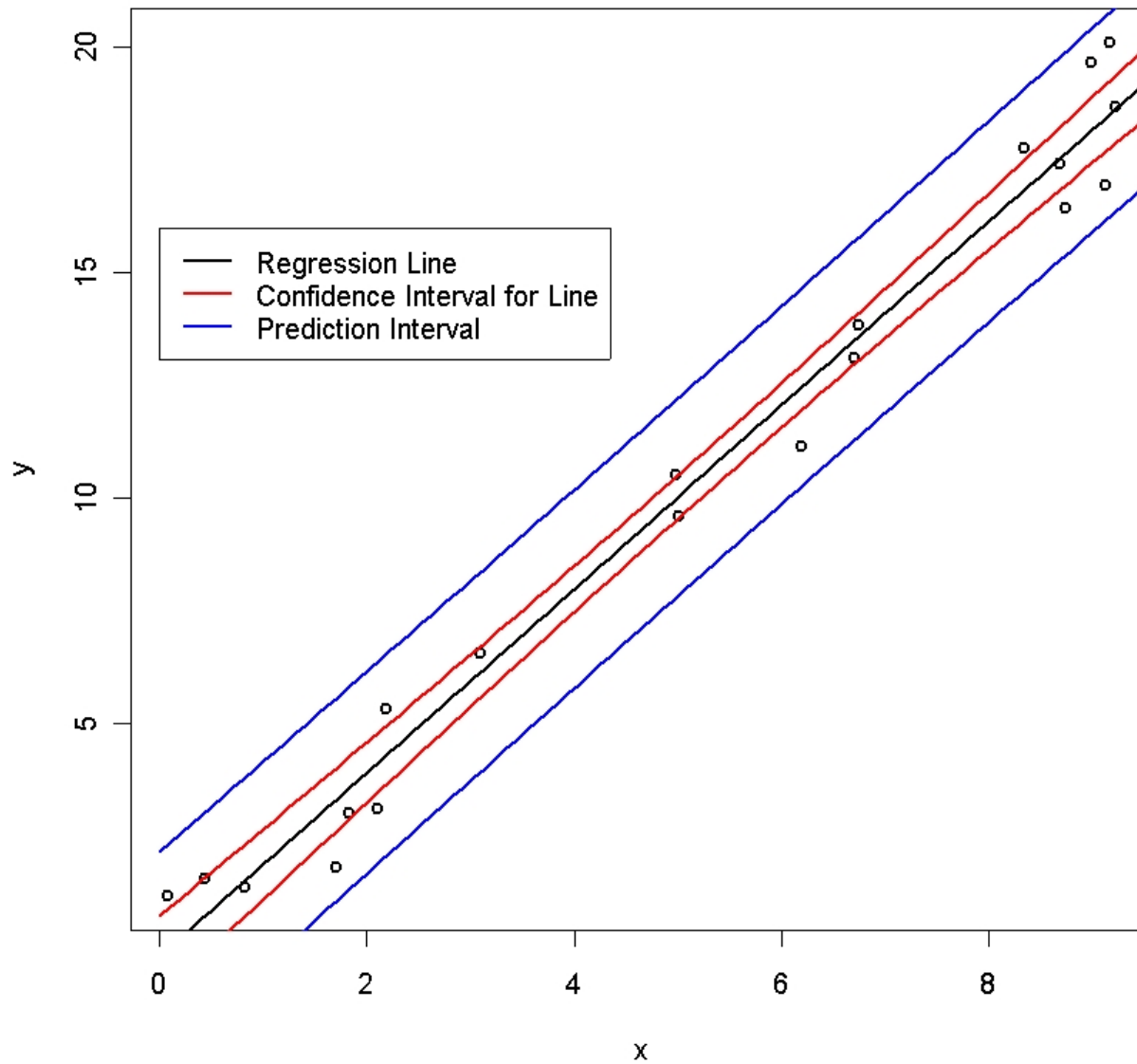
$$x_{i+1} = 0.95x_i + \varepsilon_i$$

$$y_{i+1} = 0.95y_i + \eta_i$$

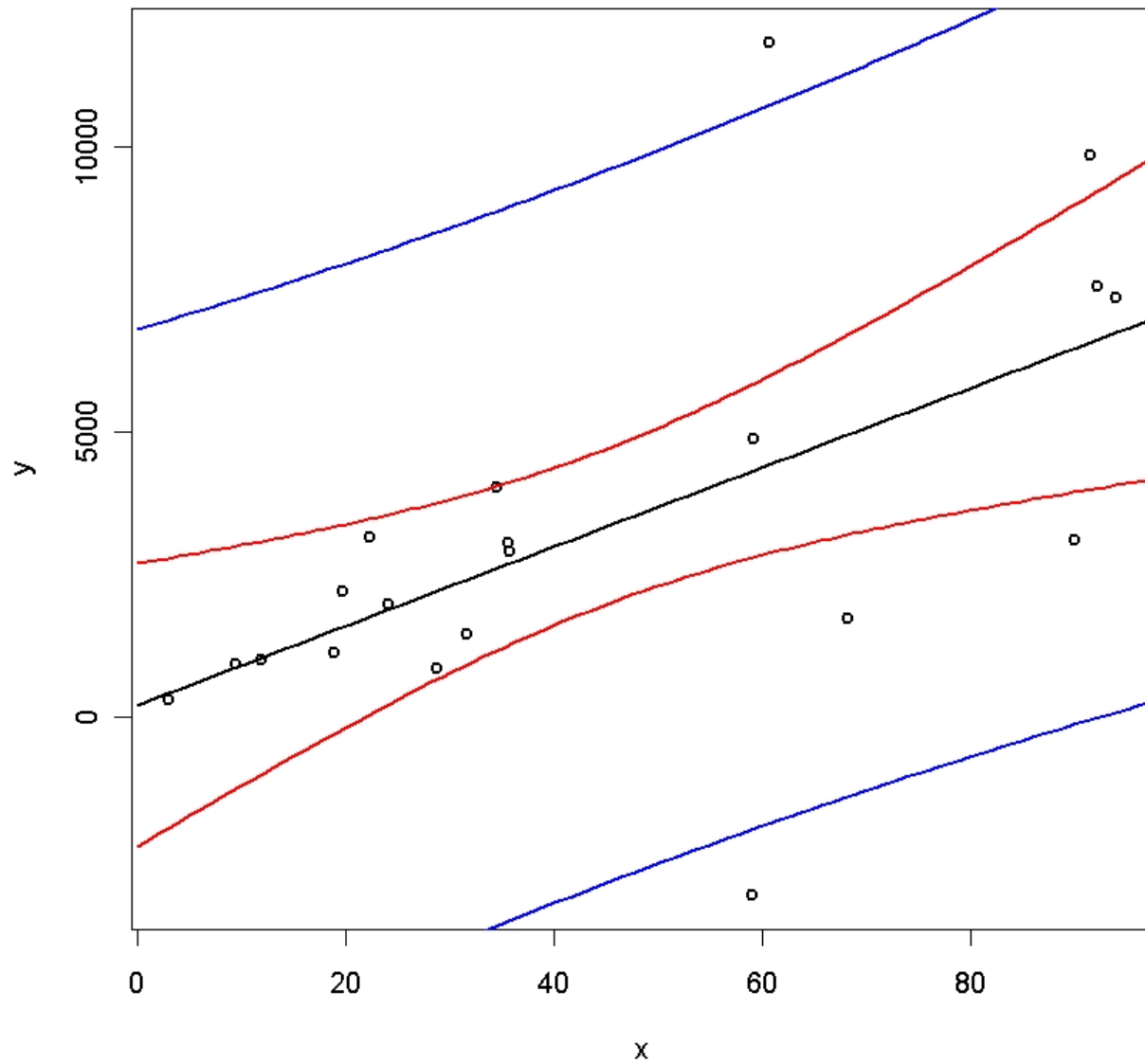
Constant Variance

- Constant variance, or homoscedasticity, means that the variability is the same in all parts of the prediction function
- If this is not the case, the predictions may be on the average correct, but the uncertainties associated with the predictions will be wrong
- Heteroscedasticity is non-constant variance

Confidence and Prediction Limits



Confidence and Prediction Limits



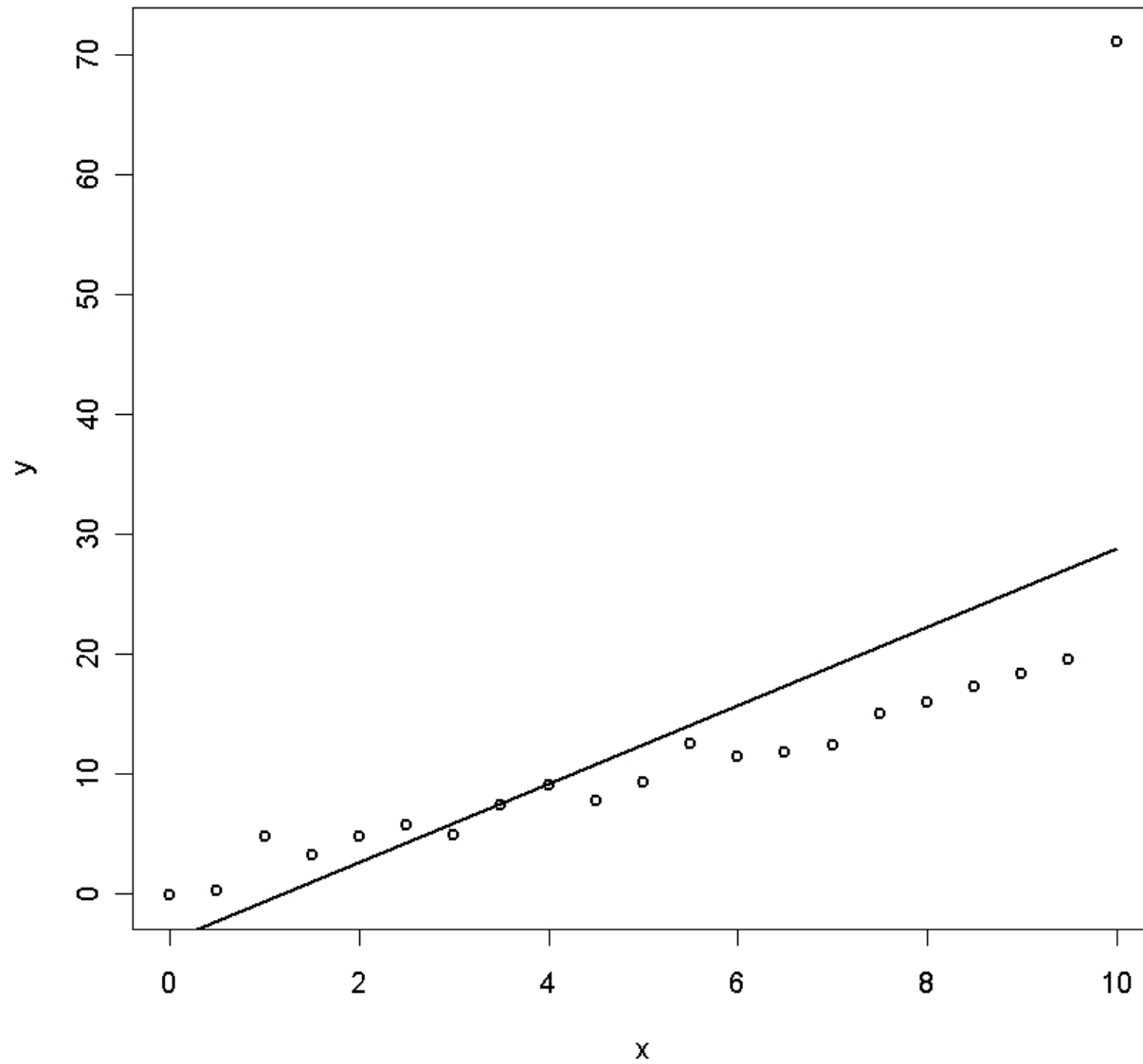
Consequences of Heteroscedasticity

- Predictions may be unbiased (correct on the average)
- Prediction uncertainties are not correct; too small sometimes, too large others
- Inferences are incorrect (is there any relationship or is it random?)

Normality of Errors

- Mostly this is not particularly important
- Very large outliers can be problematic
- Graphing data often helps
- If in a gene expression array experiment, we do 40,000 regressions, graphical analysis is not possible
- Significant relationships should be examined in detail

Consequences of Outliers



“Assumptions” and Actions

- Linearity is most important. If the relationship is non-linear there are a number of possible solutions:
 - Perhaps the relationship may be more nearly linear on the log scale. We may need to take logs of y or x or both.
 - Sometimes we use other transformations than the log, such as the square root or reciprocal.
 - Sometimes instead we fit a non-linear function directly such as a polynomial or log-logistic curve, depending on the application. We will be able to do the first by the end of the quarter.

Linear Regression

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

Quadratic Regression (Curved)

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

Nonlinear Regression Transformed to Linearity

$$y = \alpha_0 e^{\beta_1 x} + \epsilon$$

$$\log(y) = \log(\alpha_0) + \beta_1 x + \epsilon$$

Intrinsically Nonlinear Regression

$$y = \beta_0 + e^{\beta_1 x} + \epsilon$$

Independence

- This is also a highly important assumption.
- A common way in which it may fail is if the x and y values are sequential (as in annual data, 1960–2010).
- You should assume that time-series data are not statistically independent. There are tests and methods for this situation.
- Another way that this can happen is if the data are collected in batches; for example, if 16 observations are collected on four four-well microplates. This can be fixed by adding a variable for which plate the observation was collected on.

Constant Variance

- A small degree of non-constant variance is not really of concern.
- If the variance rises with the mean, so that the coefficient of variance (standard deviation divided by the mean) is roughly constant, then on the log scale the variance is roughly constant.
- If this method cannot fix the problem, and if the problem is large, meaning highly visible on plots, then a different transformation or else weighting can be used.

Normality

- In itself, this is not important.
- The main issue is with large outliers, and in simple linear regression you can see this on a plot of the data or on a plot of residuals vs. fitted values.
- You should always look for outliers in plots and investigate the possible reasons.
- Otherwise, you should not be excessively concerned with whether the distribution is normal.

Inference on Coefficients

```
>> wrightmdl = fitlm(stdwright,miniwright)
```

```
wrightmdl =
```

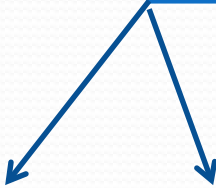
```
Linear regression model:
```

```
y ~ 1 + x1
```

```
Estimated Coefficients:
```

	Estimate	SE	tStat	pValue
(Intercept)	39.34	38.704	1.0164	0.32554
x1	0.91735	0.083365	11.004	1.3995e-08

Tests that the coefficient
is zero.



```
Number of observations: 17, Error degrees of freedom: 15
```

```
Root Mean Squared Error: 38.8
```

```
R-squared: 0.89, Adjusted R-Squared 0.882
```

```
F-statistic vs. constant model: 121, p-value = 1.4e-08
```

The statistical model is that the expected value of Y is a linear function of X

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

$$\epsilon \sim N(0, \sigma^2)$$

The errors are statistically independent, all with the distribution $N(0, \sigma^2)$, so

$$E(Y) = \beta_0 + \beta_1 E(X)$$

If we have a sample $(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)$ and we fit the least squares line we get

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

$\hat{\beta}_0$ is an estimate of β_0 and $\hat{\beta}_1$ is an estimate of β_1 , each with its standard error.

The standard errors of these estimates depend on the standard deviation of errors from the regression line.

The sum of squares of errors SSE is $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ and this has degrees of freedom $n - 2$

because we have fit two coefficients to the original n numbers.

$$s^2 = \text{MSE} = (n - 2)^{-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

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

$$\epsilon \sim N(0, \sigma^2)$$

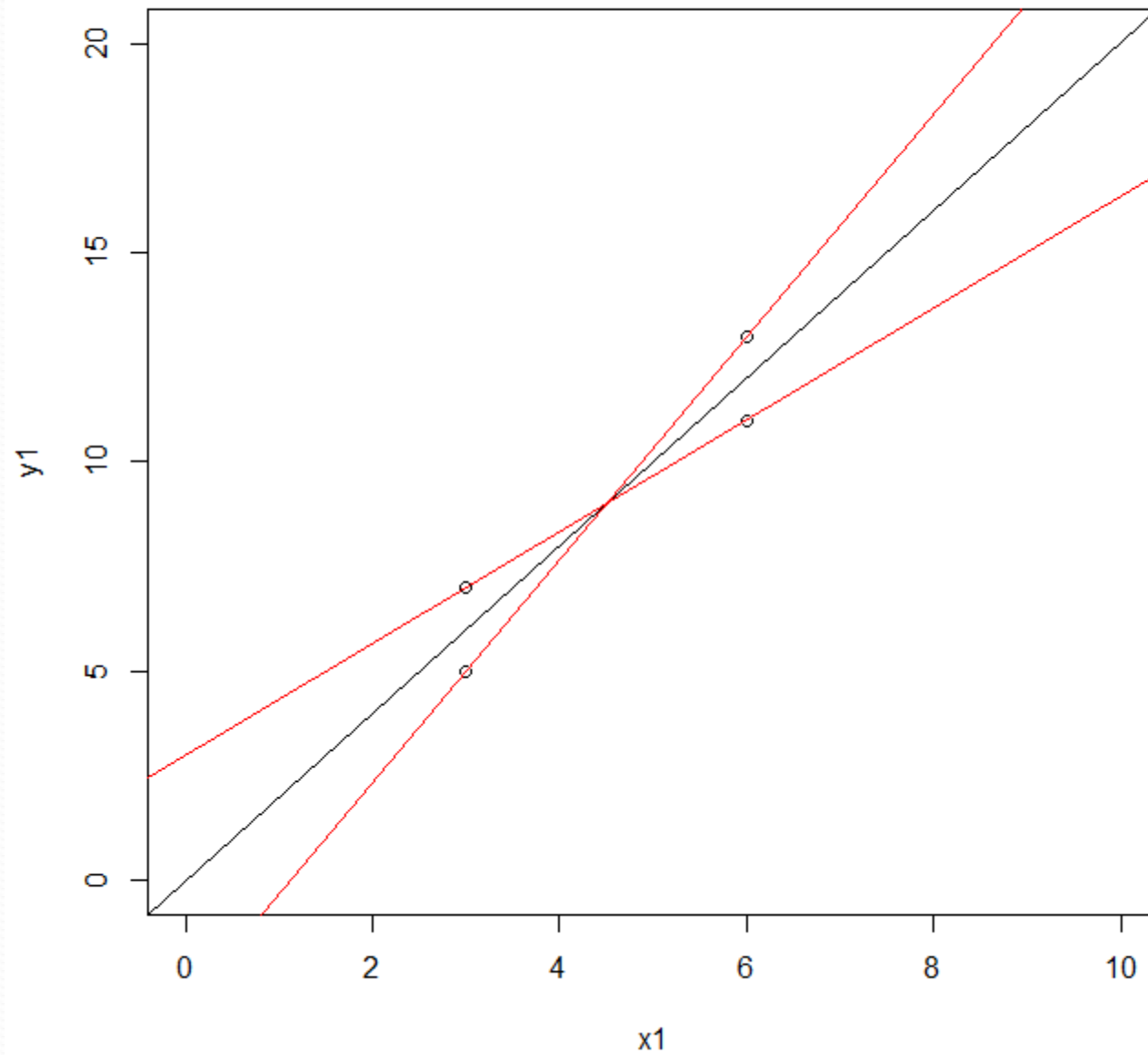
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

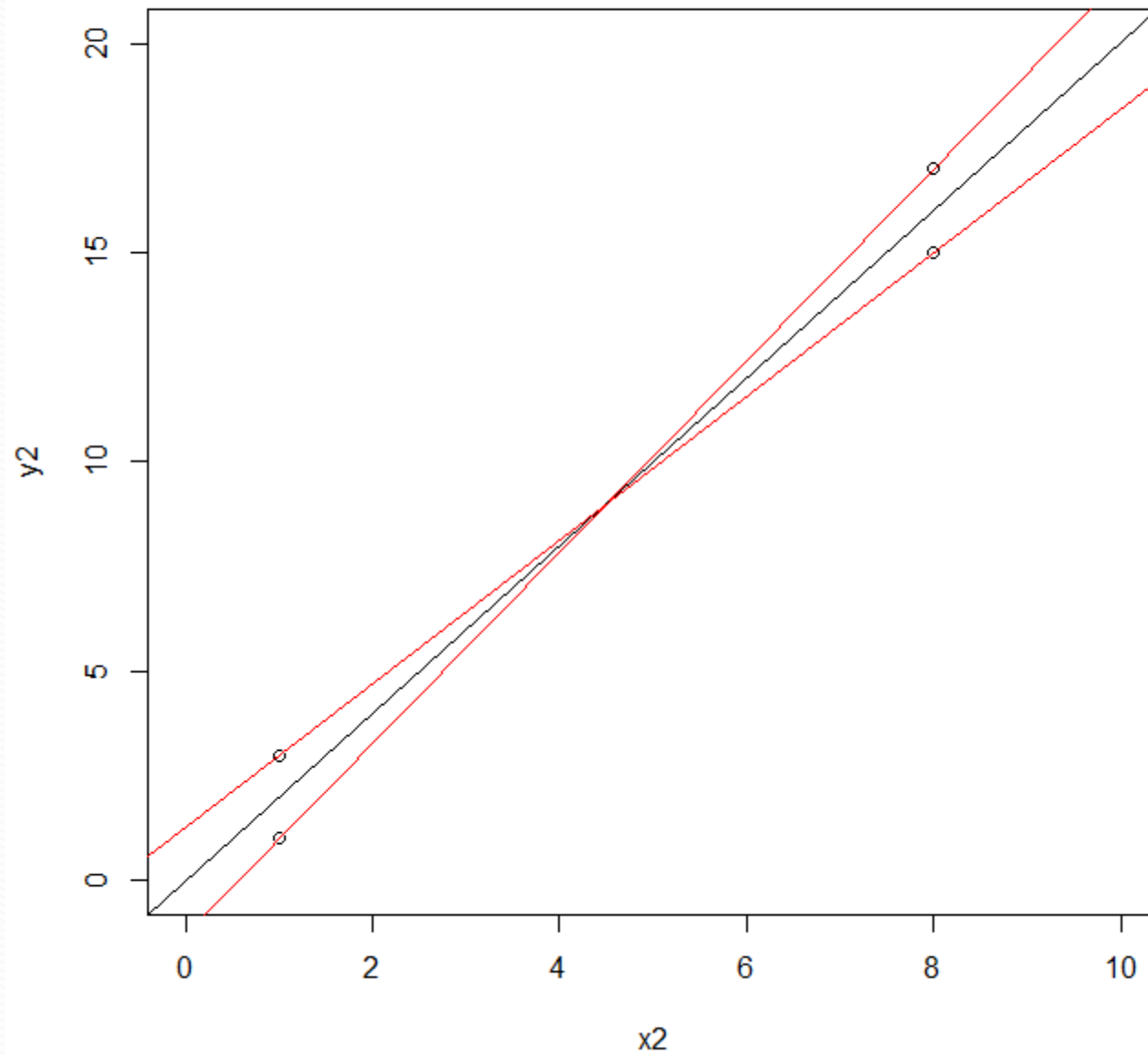
$$s^2 = \text{MSE} = (n-2)^{-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

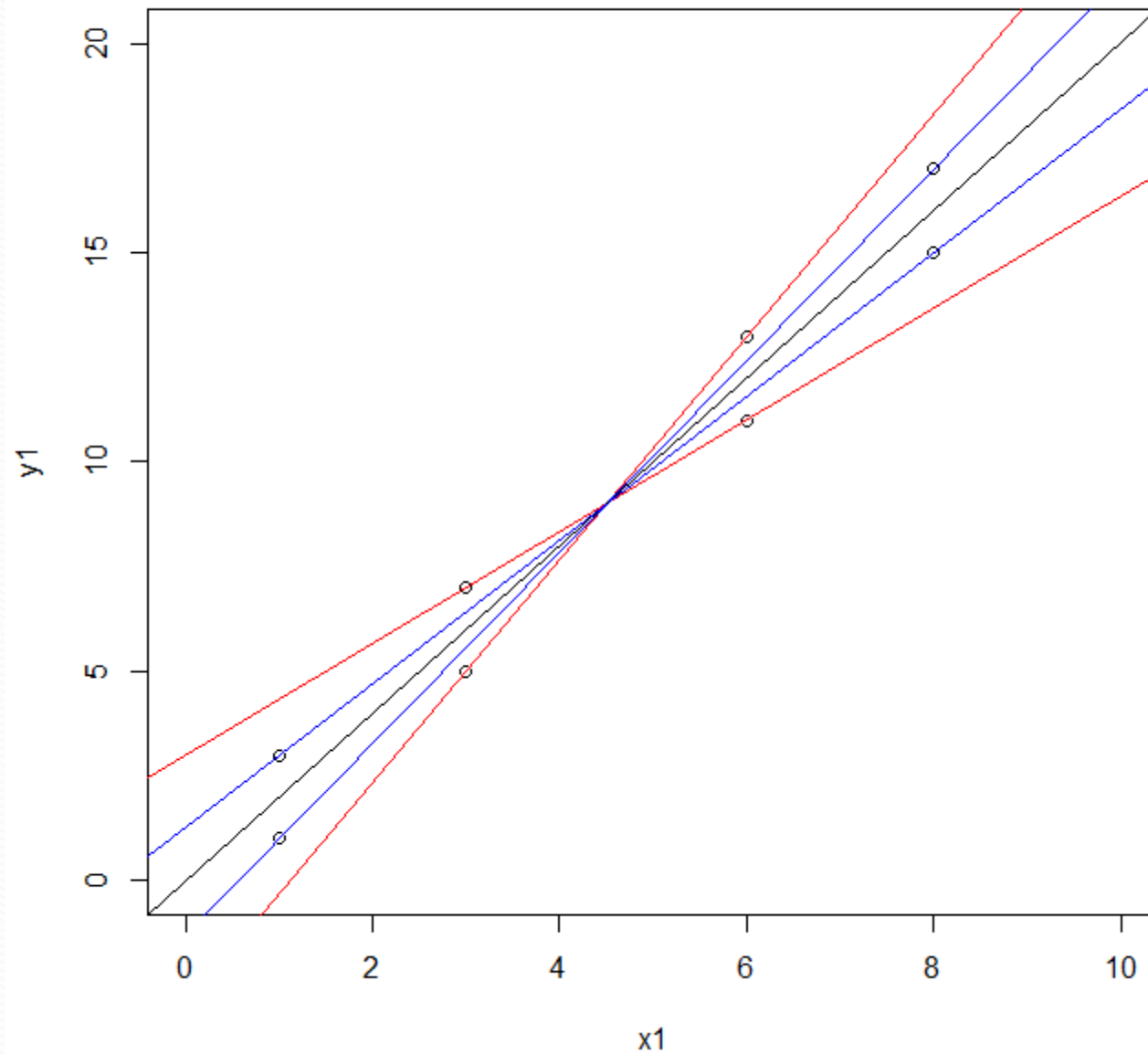
The larger s is, the more uncertain the estimates are.

$$s_{\hat{\beta}_1}^2 = \frac{s^2}{\text{SSX}} \text{ where } \text{SSX} = \sum_{i=1}^n (x_i - \bar{x})^2$$

So the more spread out the x 's are, the less error in the slope (and also intercept).







Inferences on the Slope and Intercept

- We will usually obtain the standard errors from a computer analysis.
- The hypothesis test the coefficient is 0 is usually given.
- For the slope, this is a test of association.
- For the intercept it may be of little interest.
- A confidence interval is found using the t percentage point and the given standard error.

	Estimate	SE	tStat	pValue
(Intercept)	39.34	38.704	1.0164	0.32554
x1	0.91735	0.083365	11.004	1.3995e-08

Number of observations: 17, Error degrees of freedom: 15

Root Mean Squared Error: 38.8

R-squared: 0.89, Adjusted R-Squared 0.882

F-statistic vs. constant model: 121, p-value = 1.4e-08

In this case, when both x and y are meant to measure the same thing, we might be especially interested in the hypotheses that the slope is 1 and the intercept is 0.

In that case, the model $y = x + \epsilon$ may be tenable. The first line gives a test that the intercept is 0, and it is not rejected. The second line shows that the slope is not 0, but we want to know if it is 1. With 15df, the t statistic for 95% confidence is 2.132. $0.9173 \pm (2.132)(0.08337) = 0.9173 \pm 0.1777$ or (0.7396, 1.095).

The test that $\beta_1 = 1$ is given by

$$\frac{0.9173 - 1}{0.08337} = \frac{-0.0827}{0.08337} = -0.9920 = t_{15} \text{ so } p = 0.34$$

Inference on a Mean Response

The predicted value \hat{y} for a particular value of the predictor x is

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

Its uncertainty is a combination of the uncertainties of the two estimated coefficients:

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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1(x - \bar{x}) + \hat{\beta}_1 \bar{x} = \hat{\beta}_0 + \hat{\beta}_1(x - \bar{x}) + \bar{y} - \hat{\beta}_0 = \bar{y} + \hat{\beta}_1(x - \bar{x})$$

It turns out that \bar{y} and $\hat{\beta}_1$ are statistically independent

so the variance of the sum is the sum of the variances

$$V(\bar{y}) = \sigma^2 / n \text{ estimated by } s^2 / n$$

$$V(\hat{\beta}_1) = \sigma^2 / SSX \text{ where } SSX = \sum_{i=1}^n (x_i - \bar{x})^2 \text{ estimated by } s^2 / SSX$$

$$V(\hat{\beta}_1(x - \bar{x})) = (x - \bar{x})^2 \sigma^2 / SSX \text{ estimated by } (x - \bar{x})^2 s^2 / SSX$$

$$V(\hat{y}) = s^2 \left[1/n + (x - \bar{x})^2 / SSX \right]$$

Inference on a Mean Response

$$V(\hat{y}) = s^2 \left[1/n + (x - \bar{x})^2 / SSX \right] \text{ where } SSX = \sum_{i=1}^n (x_i - \bar{x})^2$$

Usually we get the predictions and uncertainties from MATLAB

```
>> [ypred yci] = predict(wrightmdl)
ypred =
492.5101
401.6927
512.6918
437.4693
475.9979
550.3030
418.2049
444.8080
635.6164
436.5519
421.8743
641.1205
284.2722
477.8326
202.6282
427.3784
431.0478
```



```
>> [stdwright miniwright ypred yci]
```

494.0000	512.0000	492.5101	471.0120	514.0082
395.0000	430.0000	401.6927	379.3598	424.0256
516.0000	520.0000	512.6918	489.4951	535.8885
434.0000	428.0000	437.4693	417.2094	457.7291
476.0000	500.0000	475.9979	455.4361	496.5596
557.0000	600.0000	550.3030	522.7146	577.8915
413.0000	364.0000	418.2049	397.0845	439.3254
442.0000	380.0000	444.8080	424.7028	464.9133
650.0000	658.0000	635.6164	594.8671	676.3657
433.0000	445.0000	436.5519	416.2658	456.8380
417.0000	432.0000	421.8743	400.9664	442.7823
656.0000	626.0000	641.1205	599.4398	682.8012
267.0000	260.0000	284.2722	246.0169	322.5274
478.0000	477.0000	477.8326	457.1891	498.4760
178.0000	259.0000	202.6282	150.2449	255.0115
423.0000	350.0000	427.3784	406.7473	448.0095
427.0000	451.0000	431.0478	410.5725	451.5231

We have a data point where $x = \text{stdwright} = 494$ and $y = \text{miniwright} = 512$. When $x = 494$, the predicted value of y is 492.5 with CI (471.0, 514.0)

```
>> wrightmdl = fitlm(stdwright,miniwright)
```

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	39.34	38.704	1.0164	0.32554
x1	0.91735	0.083365	11.004	1.3995e-08

Number of observations: 17, Error degrees of freedom: **15**

Root Mean Squared Error: **38.8**

```
>> var(stdwright)*16
```

2.1646e+05

```
>> mean(stdwright)
```

450.3529

If $x = \text{stdwright} = 494$, then the variance of the prediction 512 is $(38.8)^2 [1/17 + (494 - 450.3529)^2/2.1646e+05] = 101.8048$ and the se is 10.0860

$t(.025,15) = 2.1314$ so the CI is $492.5101 \pm (2.1314)(10.0860)$ or (471.012, 514.008)

Prediction Intervals for Future Observations

- We just found a confidence interval for the true mean response at a particular value of the predictor.
- This is centered on the predicted value and has uncertainty depending on the uncertainties of the coefficients.
- The variance of a future predicted observation is the sum of the variance around the regression line and the variance of the predicted value.

$$V(\hat{y}) = s^2 \left[1/n + (x - \bar{x})^2 / SSX \right]$$

$$y_{new} = \beta_0 + \beta_1 x + \epsilon$$

If we predict y_{new} by \hat{y} , then the error in prediction is

$$y_{new} - \hat{y} = \beta_0 + \beta_1 x + \epsilon - (\hat{\beta}_0 + \hat{\beta}_1 x)$$

and this is the error in \hat{y} as a prediction of the true mean response plus ϵ

$$s_{pred}^2 = s^2 \left[1/n + (x - \bar{x})^2 / SSX \right] + s^2$$

$$s_{pred}^2 = s^2 \left[1 + 1/n + (x - \bar{x})^2 / SSX \right]$$

The prediction interval is wider than the confidence interval.

```
>> [yp2, yci2] = predict(wrightmdl, 'Prediction', 'observation')
>> [stdwright miniwright yp2 yci2]
```

494.0000	512.0000	492.5101	407.0906	577.9296
395.0000	430.0000	401.6927	316.0593	487.3261
516.0000	520.0000	512.6918	426.8291	598.5545
434.0000	428.0000	437.4693	352.3530	522.5855
476.0000	500.0000	475.9979	390.8092	561.1865
557.0000	600.0000	550.3030	463.1512	637.4549
413.0000	364.0000	418.2049	332.8797	503.5302
442.0000	380.0000	444.8080	359.7284	529.8877
650.0000	658.0000	635.6164	543.4490	727.7838
433.0000	445.0000	436.5519	351.4294	521.6744
417.0000	432.0000	421.8743	336.6015	507.1472
656.0000	626.0000	641.1205	548.5375	733.7034
267.0000	260.0000	284.2722	193.1800	375.3643
478.0000	477.0000	477.8326	392.6242	563.0410
178.0000	259.0000	202.6282	104.7592	300.4972
423.0000	350.0000	427.3784	342.1730	512.5838
427.0000	451.0000	431.0478	345.8800	516.2156

The prediction 492.5101 as an estimate of the mean response has CI (471, 514).

As an interval for future observations, we get (407, 578)

This is much wider (half widths 85 vs. 21)

The default for 'Prediction' is 'curve' which gives a CI for the mean response. The Option 'observation' gives a prediction interval.