# Stat 411/511

## ESTIMATING THE SLOPE AND INTERCEPT

Nov 20 2015

Charlotte Wickham

stat511.cwick.co.nz

# Quiz #4

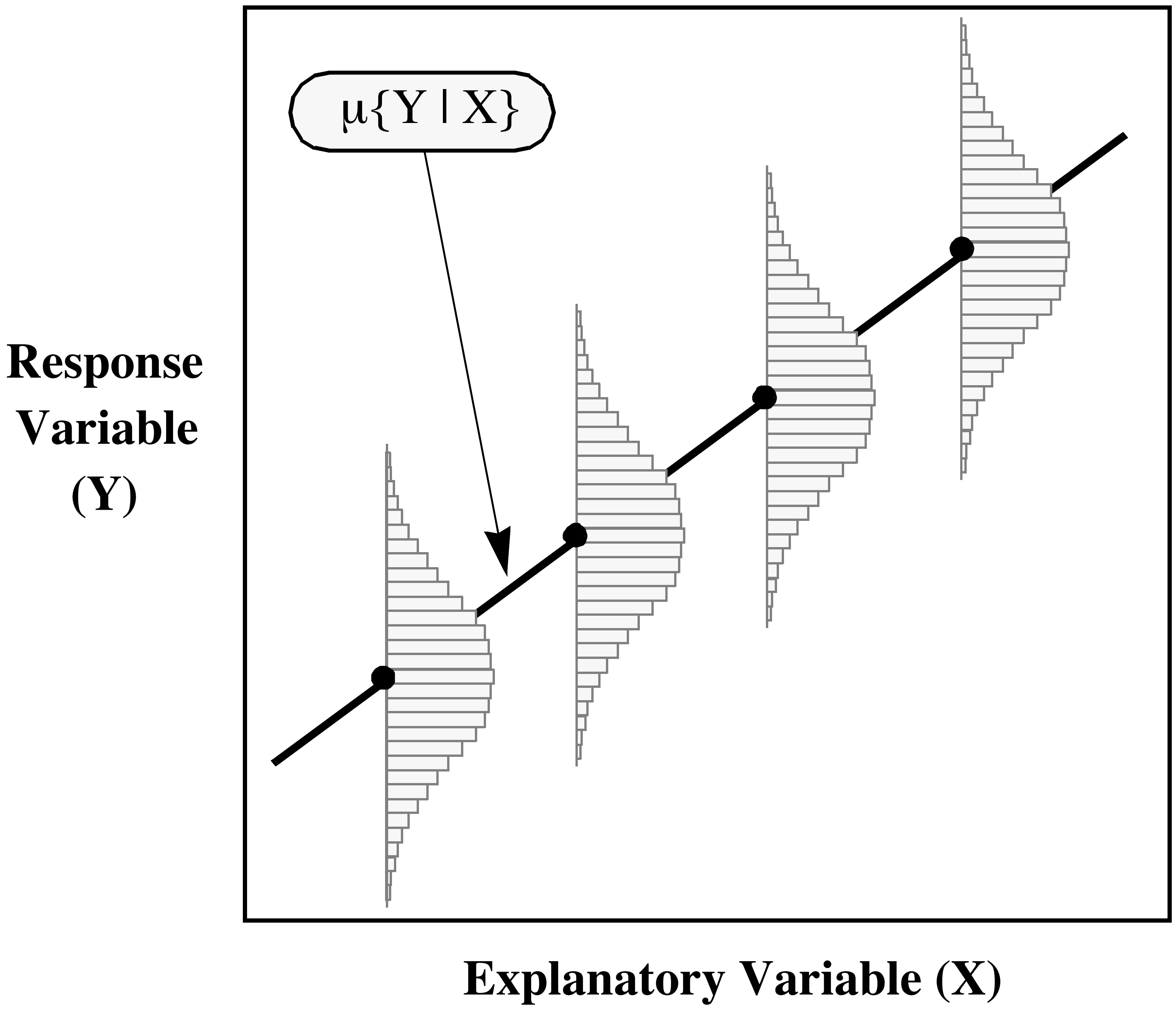
This weekend, don’t forget.

Usual format…

# Assumptions

**Display 7.5** **p. 180**

**The ideal normal, simple linear regression model**



# Assumptions

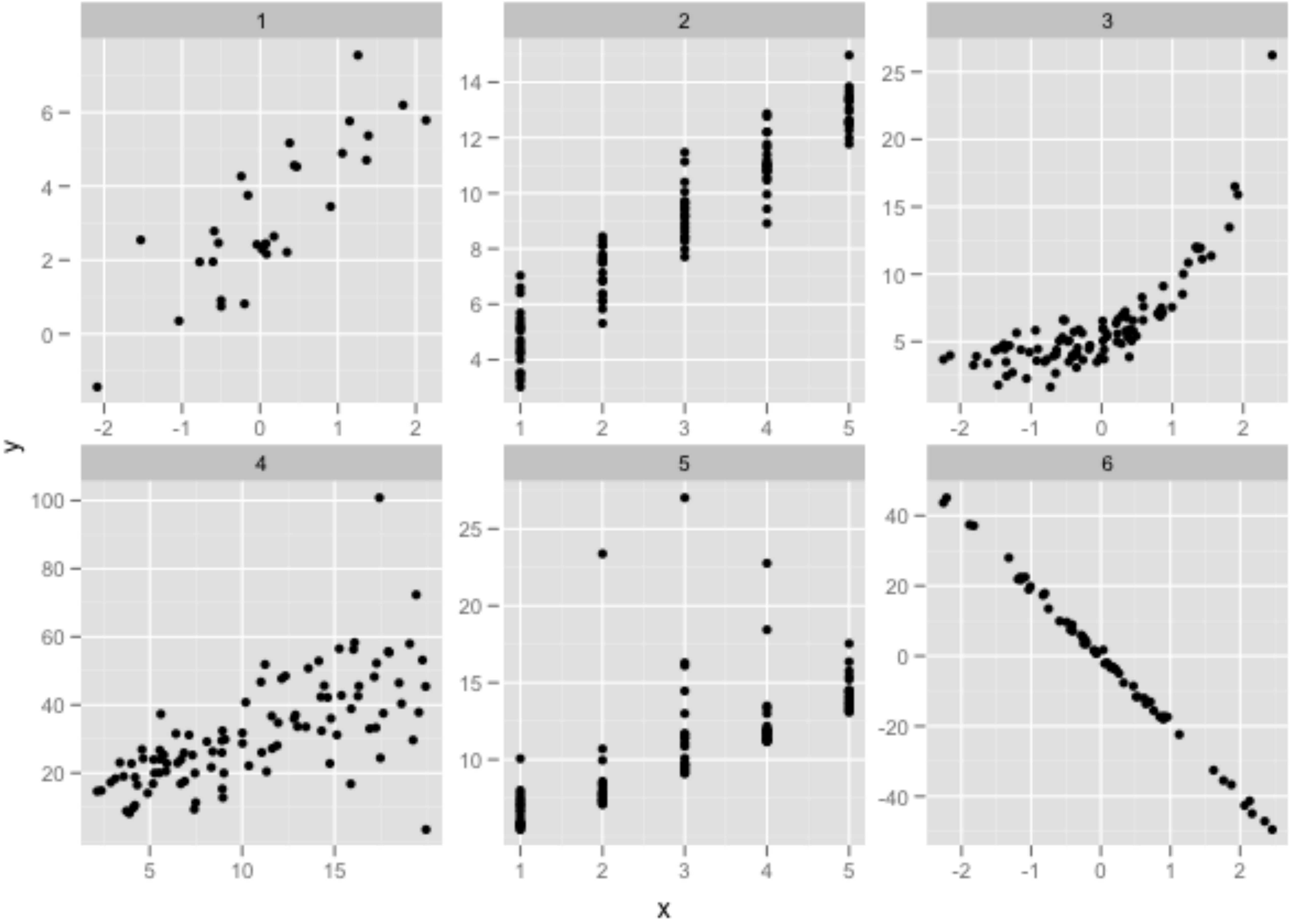
1. **Normal** subpopulation distribution of response at each value of explanatory.

2. The means of the subpopulations fall on a **straight line** function of the explanatory variable.

3. The subpopulations have the **same standard deviation**, σ. (constant spread)

4. Observations are **independent**.

# Is linear regression appropriate?



# Put a hat on it!

In statistics it is common to distinguish the **parameter** from it's **estimate** by putting a hat on it.

$\hat{\beta}_0$ is the **estimate** of the **intercept**

$\hat{\beta}_1$ is the **estimate** of the **slope**

$\hat{\mu}\{Y|X\}$ is the **estimate** of the **mean response** as a function of the explanatory variable.

# Fitted values

Once we have estimated the slope and intercept, our estimate of the **mean function** is the line defined by these estimates,

$$\hat{\mu}\{Y|X\} = \hat{\beta}_0 + \hat{\beta}_1 X$$

The **fitted value** describes the **estimated mean** for an observation. For the i$^{th}$ observation the fitted value is,

$$\text{fitted}_i = \hat{\mu}\{Y_i|X_i\} = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

# Residuals

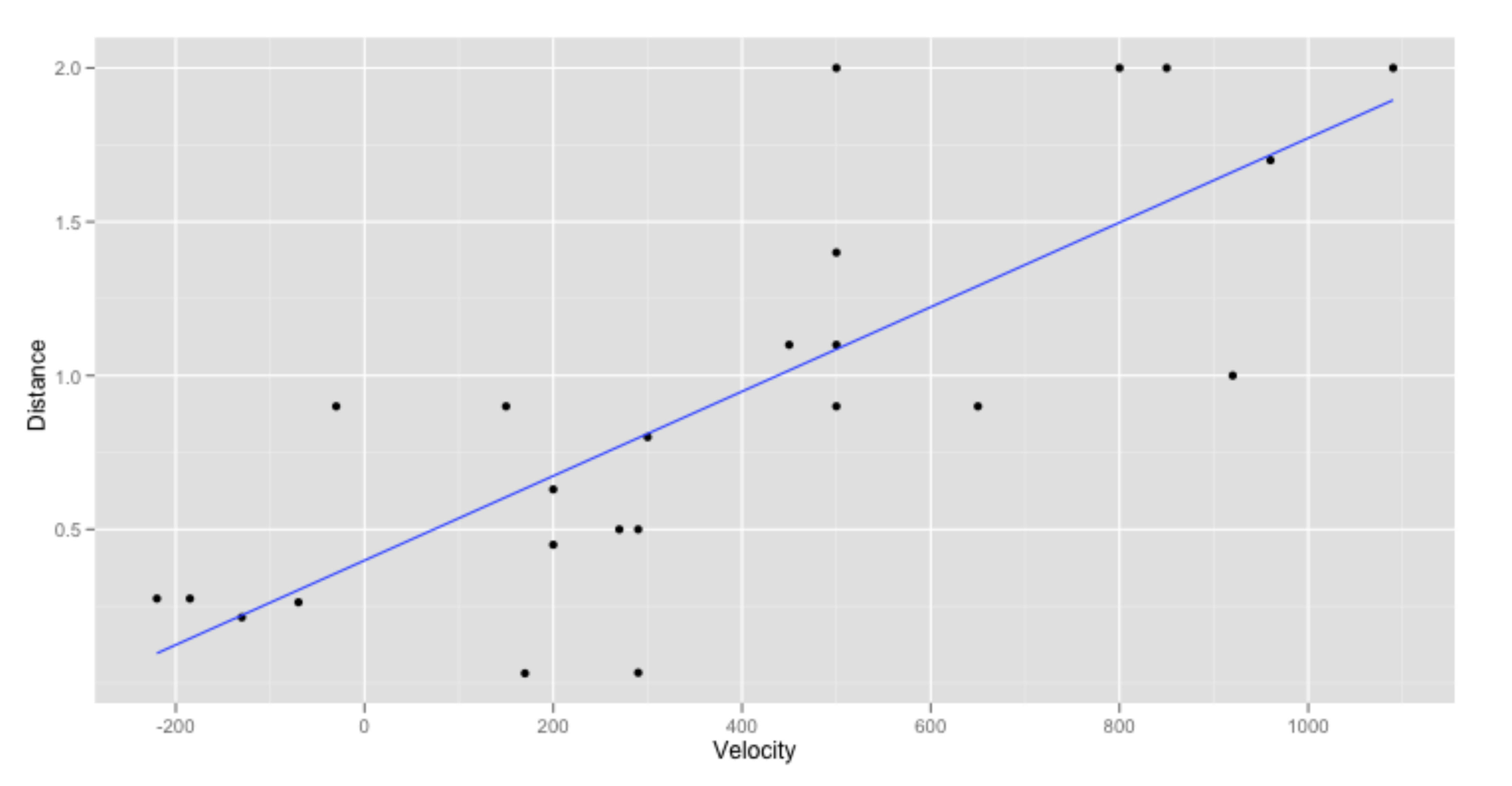
The **residual** is the difference between the observed response and it's fitted value

$$\text{residual}_i = Y_i - \text{fitted}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

This is the same definition we used in the ANOVA. The fitted value in the ANOVA case, is the group average in the full model, and overall average in the equal means model.

# Your turn

Label the fitted value for the observed nebula with a velocity of 650 km/sec and distance of 0.9 parsecs.
Draw in the residual for the same observation.



# Your turn

Calculate the fitted value, and residual for the same observation.

(velocity = 650 km/sec, distance = 0.9 parsecs)

$\hat{\beta}_0 = 0.399$ $\hat{\beta}_1 = 0.0014$

# Least squares

the line that gives the **least** possible sum of **square**d residuals

One approach to estimating the intercept and slope, is to choose the intercept and slope that **minimizes** the **sum of the squared residuals**.

It turns out this method is "optimal" in a statistical sense under our assumptions.

Of all linear unbiased estimates the least squares estimates have the lowest variance.

# Least squares estimates

The least squares estimates can be found using calculus. They are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

$$\hat{\beta}_0 = \overline{Y} - \beta_1 \overline{X}$$

$\bar{X}$ and $\bar{Y}$ are the sample averages of the explanatory and response variables

# In R

```
> lm(Distance ~ Velocity, data = case0701)

Call:
lm(formula = Distance ~ Velocity, data =
case0701)

Coefficients:
(Intercept)     Velocity
   0.399098     0.001373
```

$\hat{\beta}_0$ $\hat{\beta}_1$

# It's also easy to get residuals and fitted values in R

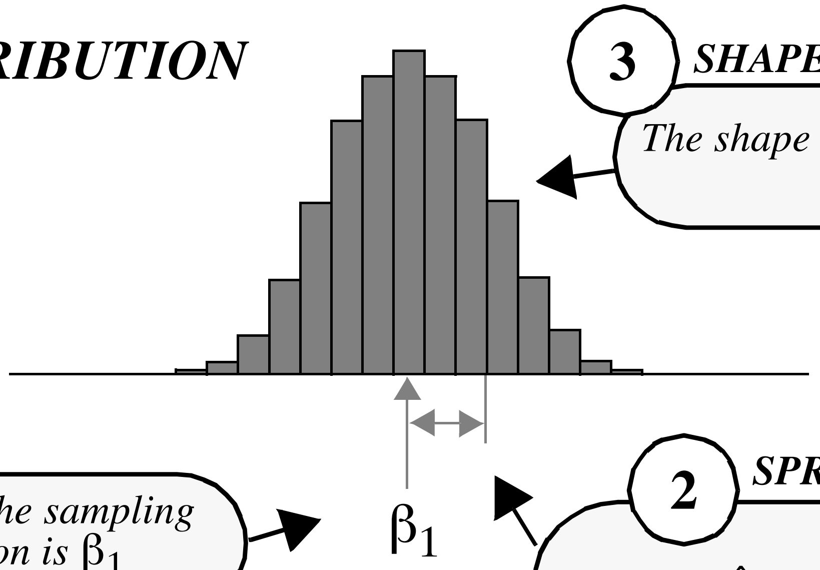
```
> fit <- lm(Distance ~ Velocity, data = case0701)
> residuals(fit)
           1            2            3            4            5            6
-0.600497351 -0.763249684 -0.006616517 -0.039992674  0.129894974  0.177947738
           7            8            9           10           11           12
-0.223685447 -0.297249686 -0.269790963 -0.043685440 -0.010979035  0.542089847
          13           14           15           16           17           18
-0.391506710  0.294961346 -0.185566293 -0.662199437  0.083080560  0.014433755
          19           20           21           22           23           24
 0.314433707 -0.017116833  0.914433731  0.433906091  0.502552897  0.104401424
> fitted(fit)
         1          2          3          4          5          6          7
0.63249735 0.79724969 0.22061652 0.30299269 0.14510503 0.09705227 0.67368544
         8          9         10         11         12         13         14
0.79724969 0.76979096 0.67368544 0.81097905 0.35791013 1.29150669 0.60503863
        15         16         17         18         19         20         21
1.08556627 1.66219944 1.01691946 1.08556627 1.08556627 1.71711688 1.08556627
        22         23         24
1.56609391 1.49744710 1.89559858
>
```

**Display 7.7** **p. 183**

**Facts about the sampling distributions of the least squares estimates of slope and intercept in the ideal normal model (from statistical theory)**

***SAMPLING DISTRIBUTION OF*** $\hat{\beta}_1$

slope



**3** ***SHAPE***

*The shape of the sampling distribution is normal*

**1** ***CENTER***

*The mean of the sampling distribution is* $\beta_1$

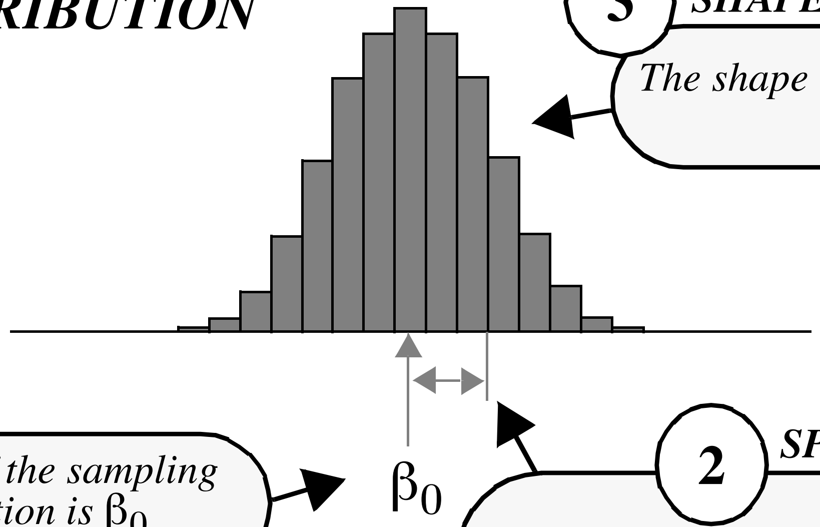
**2** ***SPREAD***

$$SD(\hat{\beta}_1) = \sigma\sqrt{\frac{1}{(n-1)s_X^2}}$$

*where* $s_X{}^2$ *is the sample variance of the X's*

***SAMPLING DISTRIBUTION OF*** $\hat{\beta}_0$

intercept



**3** ***SHAPE***

*The shape of the sampling distribution is normal*

**1** ***CENTER***

*The mean of the sampling distribution is* $\beta_0$

**2** ***SPREAD***

$$SD(\hat{\beta}_0) = \sigma\sqrt{\frac{1}{n}+\frac{\bar{X}^2}{(n-1)s_X^2}}$$

Normal

Centered around their respective parameter

(also known as unbiased)

SD depends on σ and the variation in the explanatory variable. SD gets smaller with bigger samples.

$s_X$ = sample standard deviation of X's

# Need to estimate σ

Remember **σ** is the standard deviation of the subpopulation at each value of the explanatory variable (a parameter). It measures the variation of the response around it’s mean.

The **residuals** provide an **estimate** of this variation,

$$\hat{\sigma} = \sqrt{\frac{\text{Sum of squared residuals}}{\text{Degrees of freedom}}}$$

# Degrees of freedom

**General rule**

Degrees of freedom associated with a set of residuals is the **number of observations** minus the **number of parameters for the mean.**

In simple linear regression:

$$\hat{\mu}\{Y|X\} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Two parameters

Degrees of freedom = n - 2

# Need to estimate σ

Remember σ is the standard deviation of the subpopulation at each value of the explanatory variable.

The residuals provide a measure of this variation,

$$\hat{\sigma} = \sqrt{\frac{\text{Sum of squared residuals}}{\text{Degrees of freedom}}}$$

$$\hat{\sigma} = \sqrt{\frac{\text{Sum of squared residuals}}{n-2}}$$

```
> sum_sq_residuals <- sum(residuals(fit)^2)
> df <- length(residuals(fit)) - 2
> df
[1] 22
> sqrt(sum_sq_residuals/df)
[1] 0.4049588
```

OR

```
> summary(fit)

Call:
lm(formula = Distance ~ Velocity, data = case0701)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.7632 -0.2352 -0.0088  0.2072  0.9144 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.3990982  0.1184697   3.369  0.00277 ** 
Velocity    0.0013729  0.0002274   6.036 4.48e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.405 on 22 degrees of freedom
Multiple R-squared: 0.6235,	Adjusted R-squared: 0.6064 
F-statistic: 36.44 on 1 and 22 DF,  p-value: 4.477e-06
```

# Standard errors

Plug in estimates into formula for standard deviations

$$\mathrm{SE}_{\hat{\beta}_1} = \hat{\sigma}\sqrt{\frac{1}{(n-1)s_X^2}}$$

d.f. = n - 2

$$\mathrm{SE}_{\hat{\beta}_0} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{(n-1)s_X^2}}$$

```
> summary(fit)

Call:
lm(formula = Distance ~ Velocity, data = case0701)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.7632 -0.2352 -0.0088  0.2072  0.9144 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.3990982  0.1184697   3.369  0.00277 ** 
Velocity    0.0013729  0.0002274   6.036 4.48e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.405 on 22 degrees of freedom
Multiple R-squared: 0.6235,	Adjusted R-squared: 0.6064 
F-statistic: 36.44 on 1 and 22 DF,  p-value: 4.477e-06
```

# Your turn

We are less certain about our estimate of the slope when we have a larger standard error.

$$\mathrm{SE}_{\hat{\beta}_1} = \hat{\sigma}\sqrt{\frac{1}{(n-1)s_X^2}}$$

Will we be less or more certain about the slope for:

larger sample size, n?

larger subpopulation variation, σ?

larger variation in observed explanatory values, $s_x$?

# Three types of inference

Inference on the slope or intercept

uncertainty comes from sampling error in a single parameter

Inference about the mean response
(at a given explanatory value)

uncertainty comes from sampling error in both parameters

Prediction of a new response
(at a given explanatory value)

uncertainty comes from sampling error in both parameters
and variability in subpopulations