Interactions I

Add a predictor involving the product of two existing predictors:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} \cdot X_{i2} + \epsilon_i.$$

β_3 is called the interaction term.

Important: Now Y is no longer a linear function of X_1, X_2, it is a linear function of $X_1, X_2, X_1 \cdot X_2$.

When X_1 increases with X_2 fixed, Y increase depends on X_2, it is $\beta_1 + \beta_3 X_2$.

Interactions II

Back to cats:If one of the variables is an indicator the interaction produces a separate slope as well as having a separate intercept:
Lines go through points of means of each sub-population

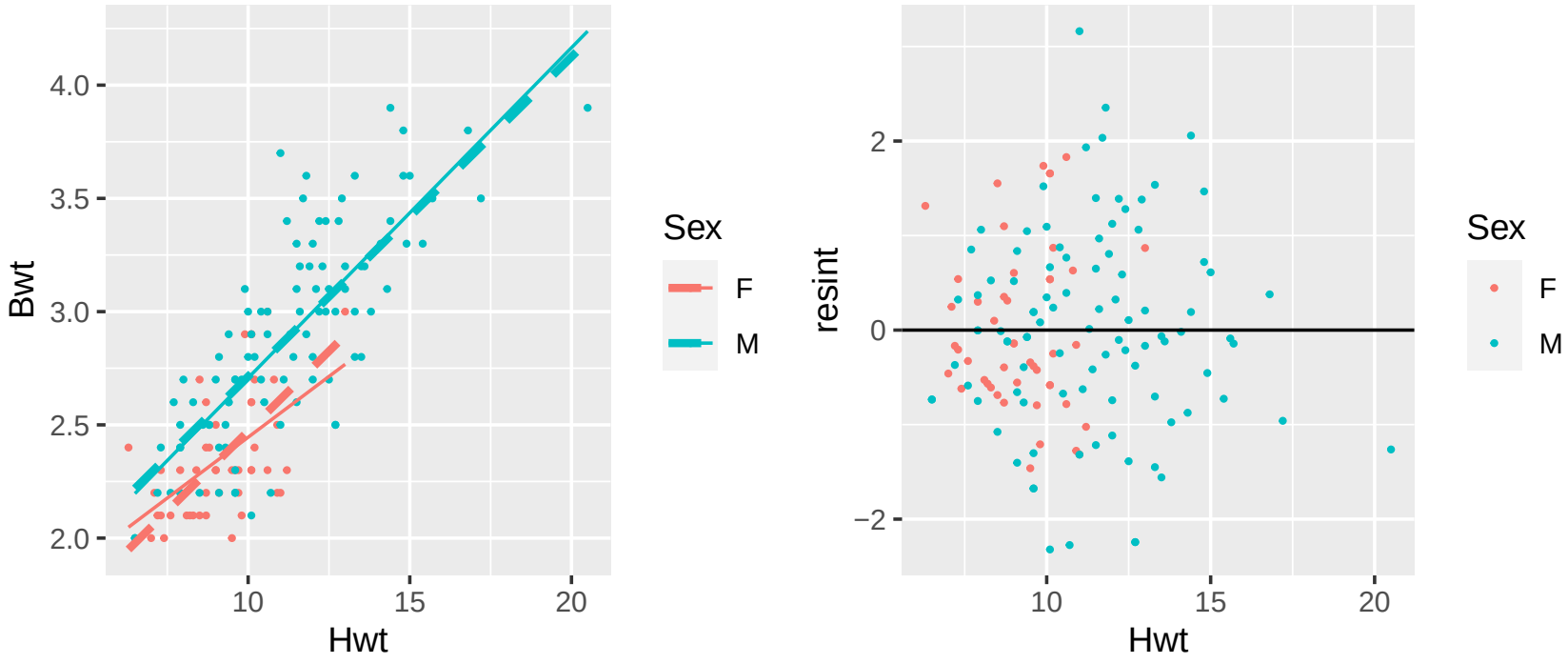
```
mod=lm(Bwt~Hwt+Sex,data=cats)
modint=lm(Bwt~Hwt+Sex+Hwt:Sex,data=cats)
print(summary(modint)$coefficients)
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  1.37152432 0.27342241  5.0161372 1.571750e-06
## Hwt          0.10737192 0.02940142  3.6519303 3.668392e-04
## SexM        -0.12265323 0.30109805 -0.4073531 6.843708e-01
## Hwt:SexM     0.03845299 0.03134602  1.2267264 2.219845e-01
```

If Female: `Bwt = 1.3715 + 0.1074*Hwt`

If Male: `Bwt = 1.25 + 0.146* Hwt`

Interactions III



This isn't the same as estimating a separate regression for each sex because we are using a common variance - we are pooling data for estimating the variance.

Polynomials I

Sometimes we want to add polynomials in the original predictors.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i.$$

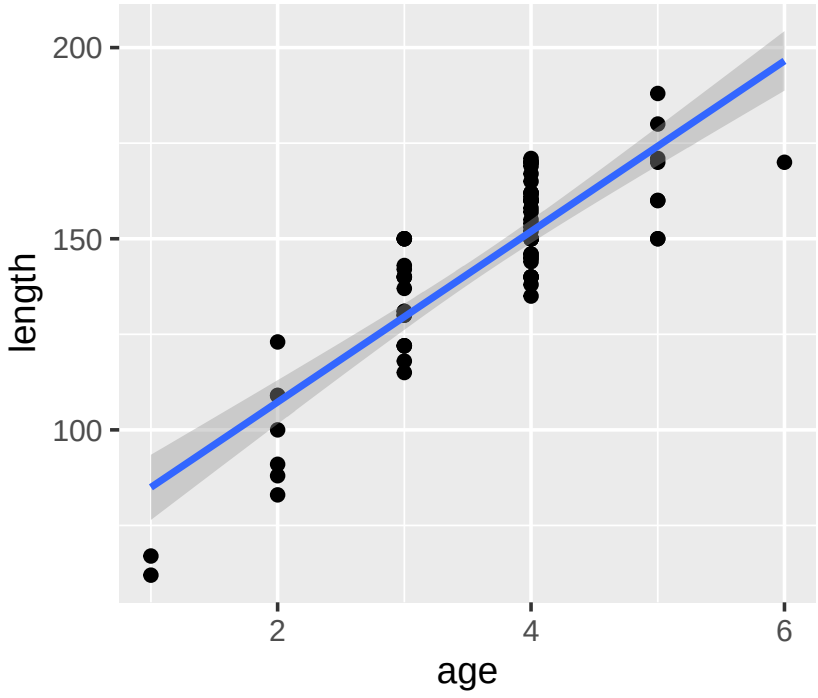
Again, the response is no longer linear in X, rather it is linear in X and X^2.

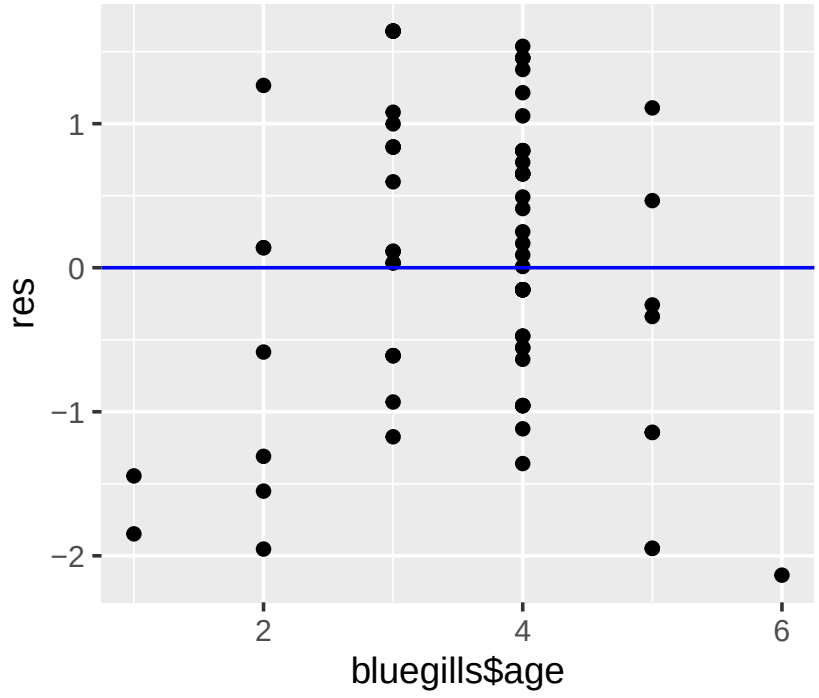
Polynomials II

Rleationship between age and length of bluegills:

```
bluegills=read.table("bluegills.txt",header = TRUE)
mod=lm(length~age,data=bluegills)
p1=qplot(age,length,data=bluegills,geom=c("point","smooth"),method=c("lm"))
res=(mod$res-mean(mod$res))/sd(mod$res)
p2=qplot(bluegills$age,res)+geom_hline(yintercept=0,col="blue")
grid.arrange(p1,p2,ncol=2)
```

Polynomials III

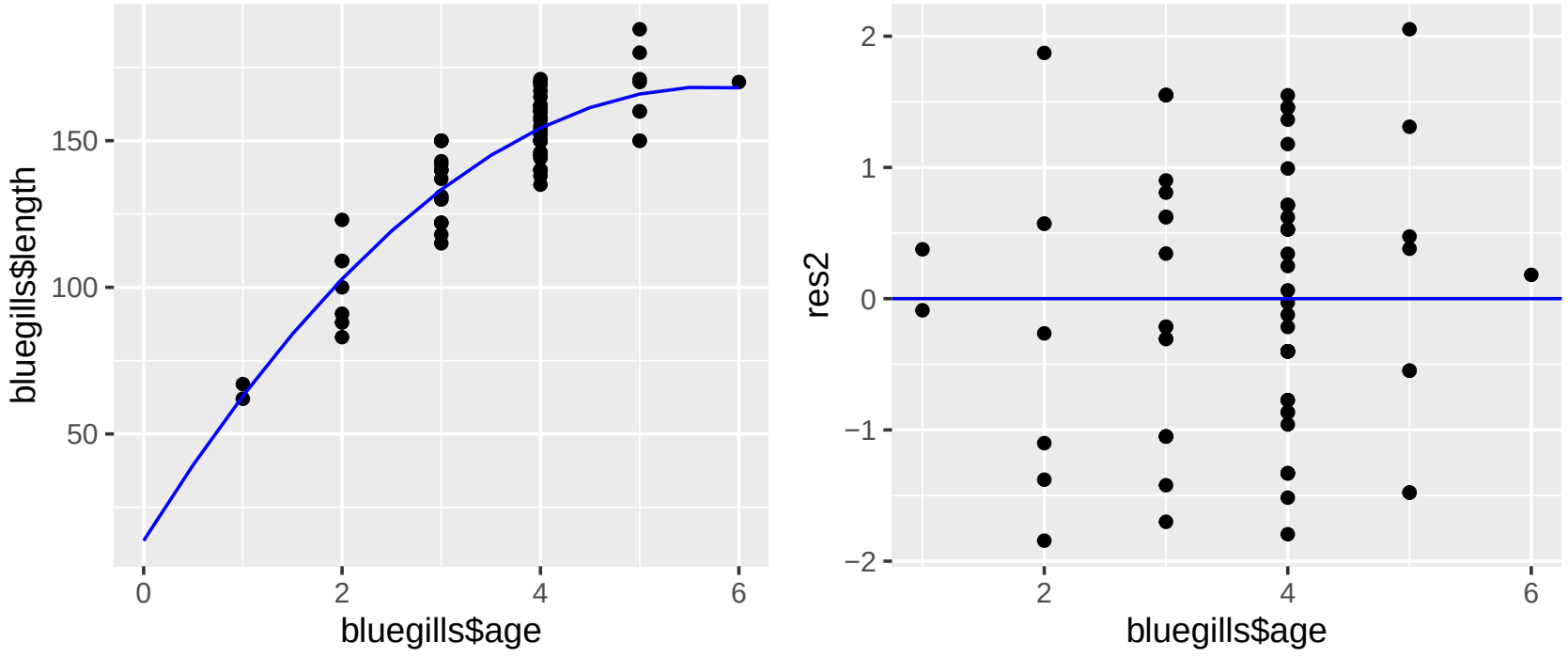




Polynomials IV

Try adding a second order term:

```
mod2=lm(length~age+I(age^2),data=bluegills)
p1=qplot(bluegills$age,bluegills$length,geom=c("point"))+geom_line(aes(x=seq(0,6,.5),y=predict(mod2,newdat
res2=(mod2$res-mean(mod2$res))/sd(mod2$res)
p2=qplot(bluegills$age,res2)+geom_hline(yintercept=0,col="blue")
grid.arrange(p1,p2,ncol=2)
```



Polynomials V

```
summary(mod2)


##
## Call:
## lm(formula = length ~ age + I(age^2), data = bluegills)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.846  -8.321  -1.137   6.698  22.098
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.622     11.016   1.237     0.22
## age           54.049      6.489   8.330 2.81e-12 ***
## I(age^2)      -4.719      0.944  -4.999 3.67e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.91 on 75 degrees of freedom
## Multiple R-squared:  0.8011,Adjusted R-squared:  0.7958
## F-statistic: 151.1 on 2 and 75 DF,  p-value: < 2.2e-16
```

Polynomials VI

Interactions and polynomials are examples of **non-linear** regression.

$$Y_i = f(X_{1i}, X_{2i}) + \epsilon_i.$$

The function f is non-linear in the original variables. But it is linear in terms of polynomials defined in terms of the original variables, e.g

$$X_{1i}^2, X_{1i} \cdot X_{2i}, ...$$

.

Review of Linear Models: One Mean I

A model for one random sample from one population

$$\underset{\text{response (random)}}{Y} = \underset{\text{deterministic model for the mean (not random)}}{\mu} + \underset{\text{error (random)}}{\epsilon}$$

Random sample $\Longrightarrow$ (nearly) independent outcomes

$$\underset{\text{data}}{Y_i} = \underset{\text{population mean}}{\mu} + \underset{\substack{\text{random error} \\ \uparrow \\ \text{errors independent since sample random}}}{\epsilon_i} \quad \text{for } i = 1, \ldots, \underset{\text{sample size}}{n}$$

Review of Linear Models: One Mean II

ϵ_i = "errors" = difference between data and model: $\epsilon_i = Y_i - \mu$.

Suppose the errors have probability distribution

$$\epsilon_i \sim N(\underset{\text{mean}}{\underset{\downarrow}{0}}, \underset{\text{variance}}{\underset{\downarrow}{\sigma}}).$$

Then, the population mean (and mean for a random individual) is

$$E(Y_i) = E(\mu + \epsilon_i) = \mu + E(\epsilon_i) = \mu + 0 = \mu$$

and variance

$$var(Y_i) = var(\mu + \epsilon_i) = var(\epsilon_i) = \sigma^2$$

since μ is a constant and for any r.v. ϵ_i, $var(a + b\epsilon_i) = b^2\, var(\epsilon_i)$

Review of Linear Models: One Mean III

This simple model (the constant μ) is a **linear model** for the mean:

$\mu_i = \mu(x_i) = \beta_0 + \beta_1 x_i$ for constants β_0 and β_1 and n fixed values $x_1 = x_2 = \ldots = x_n = 1$.

Here, $\beta_0 = \mu$ and $\beta_1 = 0$, so the x variable is irrelevant.

Review of Linear Models: One Mean IV

Estimate the model (mean) using the Least Squares Principle

What value $\widehat{\mu}$ will **minimize sum of squared errors**?

$$\begin{aligned} \text{Sum of Squared Errors} = \quad \text{SSE} \quad &= \sum_{i=1}^{n}(Y_i - \hat{\mu}_i)^2 \\ &= \sum_{i=1}^{n}(Y_i - b_0)^2 \end{aligned}$$

$\uparrow$

We already know that $b_0 = \hat{\beta}_0 = \widehat{\mu} = \overline{Y}$ minimizes the SSE

Review of Linear Models: One Mean V

Calculate an unbiased estimate the error variance: SSE/df

We've used the estimate $\widehat{\sigma^2} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$

and we know this estimate is unbiased: $E(S^2) = \sigma^2$

How is this related to the suggested estimator SSE/df ?

$$\frac{\text{SSE}}{\text{df}} = \frac{\text{SSE}}{n-1} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1} = S^2 \quad \leftarrow \text{unbiased for } \sigma^2 .$$

$\uparrow$

We subtract 1 to find the degrees of freedom (df), since exactly 1 parameter (μ) is estimated for the linear model for the mean

Review of Linear Models: One Mean VI

When the errors have a normal distribution, the standardized ("studentized") LS estimator is t_{df}-distributed

If $\epsilon_i \sim N(0, \sigma)$ then $Y_i \sim N(\mu, \sigma)$, and $SE(\widehat{\mu}) = S/\sqrt{n}$

$$t = \frac{\hat{\mu} - E(\hat{\mu})}{SE(\widehat{\mu})}$$

has a t-distribution with $n-1$ df.
$100(1-\alpha)\%$ Confidence interval for μ:
$\widehat{\mu} \pm t^* \, SE(\widehat{\mu})$
$t^* : P(T_{n-1} > t^*) = \alpha/2.$

Special case: Y - binary estimating proportion.
$\hat{\mu} = \hat{p} = \bar{Y}$,
$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Introduction to Linear Models: Two Means I

A model for two random samples from two populations

response from population j (random)		model for mean of population j (not random)		error (random)
↓		↓		↓
Y_j	$=$	μ_j	$+$	ϵ

data		population mean		random errors			
↓		↓		↓	↓		sample sizes n_1, n_2
Y_{ji}	$=$	μ_j	$+$	ϵ_{ji}	$j = 1, 2$	$i = 1, ..., n_j$	

Random samples $\Longrightarrow$ independent samples and outcomes (nearly)

$\Longrightarrow$ the random errors are independent

Introduction to Linear Models: Two Means II

This model (separate mean μ_j for each population $j = 1, 2$) is also a **linear model** for the mean response:

$\mu_j = \beta_0 + \beta_1 x_{ji}$

for constants β_0 and β_1 and $(n_1 + n_2)$ fixed values

$x_{11} = 0, x_{12} = 0, \ldots, x_{1n_1} = 0$, for data from population 1

$x_{21} = 1, x_{22} = 1, \ldots, x_{2n_2} = 1,$ for data from population 2

The x-values denote group (population) membership.

$\mu_1 = \beta_0 + \beta_1(0) = \beta_0$ and $\mu_2 = \beta_0 + \beta_1(1) = \beta_0 + \beta_1 = \mu_1 + \beta_1$.

That is, $\beta_0 = \mu_1$ and $\beta_1 = \mu_2 - \mu_1$.

So, $H_o : \mu_1 = \mu_2$ is equivalent to $H_o : \beta_1 = 0$.

Introduction to Linear Models: Two Means III

Estimate the model (means) by the Least Squares Principle

What values of $\widehat{\mu_1}$ and $\widehat{\mu_2}$ will **minimize sum of squared errors**?

$$\text{SSE} = \sum_{j=1}^{2}\sum_{i=1}^{n_j}(\ Y_{ji} \ - \ (b_0 + b_1 x_{ji}) \ )^2$$

$\uparrow$

$\widehat{\mu}_{ji} = b_0 + b_1 x_{ji}$ estimate of mean for population j, $j = 1, 2$
Here $b_j, j = 1, 2$ is the estimate for $\beta_j, j = 1, 2$.

$b_0 = \hat{\beta}_0 = \overline{Y}_1$ and $b_1 = \hat{\beta}_1 = \overline{Y}_2 - \overline{Y}_1$ minimize the SSE.

Introduction to Linear Models: Two Means IV

This model specifies that the two populations have same variance. That is, $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
An unbiased estimate for σ^2 has the form

$$S^2 = = \frac{\text{SSE}}{\text{df}} = \frac{\text{SSE}}{n_1 + n_2 - 2}$$

The degrees of freedom are $\text{df} = n_1 + n_2 - 2$ since $n_1 + n_2$ is the total sample size, and 2 parameters μ_1, μ_2 are estimated.

$$S^2 = \frac{\text{SSE}}{\text{df}} = \frac{\sum_{j=1}^{2} \sum_{i=1}^{n_j} (\ Y_{ji} - \widehat{\mu}_j\ )^2}{n_1 + n_2 - 2}$$

$$= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = s_p^2 = \text{“pooled variance”}$$

We can show that $S^2 = s_p^2$ is an unbiased estimate for $\sigma^2 : E(s_p^2) = \sigma^2$.

Introduction to Linear Models: Two Means V

$SE(b_1) = SE(\hat{\mu}_2 - \hat{\mu}_1) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$.
If $\epsilon_{ji} \sim N(0, \sigma)$, then $Y_{1i} \sim N(\mu_1, \sigma)$ and $Y_{2i} \sim N(\mu_2, \sigma)$,

$t = \frac{(\hat{\mu}_2 - \hat{\mu}_1) \; - \; (\mu_1 - \mu_2)}{SE(\hat{\mu}_2 - \hat{\mu}_1)}$
has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.
$100(1-\alpha)\%$ Confidence interval for $\mu_1 - \mu_2$:
$(\hat{\mu}_2 - \hat{\mu}_1) \; \pm \; t^* \; SE(\hat{\mu}_2 - \hat{\mu}_1), \quad t^* : P(T_{n_1+n_2-2} > t^*) = \alpha/2.$

Special case: Y_1, Y_2 bernoulli, $\hat{\mu}_i = \hat{p}_i$,
can't assume equal variances **when estimating difference of means.**
$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n2}$.
When testing $H_0 : p_1 = p_2$, used pooled estimate
$\hat{p} = \frac{\sum_{i=1}^{n_1} x_{1i} + \sum_{i=1}^{n_2} x_{2i}}{n_1+n_2}$,
$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}(1-\hat{p})}\sqrt{1/n_1 + 1/n_2}$

Introduction to Linear Models: Two Variables I

response (random)		mean depends on another variable (x) (not random)		error (random)
$\downarrow$		$\downarrow$		$\downarrow$
Y	$=$	$\mu(x)$	$+$	ϵ

(Each x can be viewed as a populaton)
We consider the x-values as "fixed" and model the probability distribution of Y "conditional" on the observed x-values.

A simple model considers the mean to be a linear function of x :
$\mu(x) = \mu_{Y|x} = E(Y|x) = \beta_0 + \beta_1 x.$

This model is called the **simple linear regression model**.

Introduction to Linear Models: Two Variables II

The linear model is $Y = \beta_0 + \beta_1 x + \epsilon$,

and we observe pairs (x_i, Y_i). For each observation:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \ , \\ & \epsilon_i \sim \text{ independent } N(0, \sigma) \qquad i = 1, 2, \ldots, n \\ \Rightarrow \quad & Y_i - (\beta_0 + \beta_1 x_i) = \epsilon_i = \text{error or "residual"} \end{aligned}$$

$$\begin{aligned} E(Y_i) &= E(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 x_i + E(\epsilon_i) \\ &= \beta_0 + \beta_1 x_i \end{aligned}$$

$$\begin{aligned} var(Y_i) &= var(\beta_0 + \beta_1 x_i + \epsilon_i) = var(\epsilon_i) \\ &= \sigma^2 \end{aligned}$$

This is a considerable assumption: constant variance for the response variable (Y) for every value of predictor variable (x).

Introduction to Linear Models: Two Variables III

Estimate for $\mu(x)$ by the "least-squares" method: minimize the **sum of squared errors** with respect to the parameters included in the model specified for the mean:

$$\text{SSE} \;=\; \sum_{i=1}^{n}(\; Y_i \;-\widehat{Y}_i\;)^2 \qquad \text{where} \qquad \widehat{Y}_i = \widehat{\mu}(x_i) = b_0 + b_1 x_i$$

LS estimates b_0 for β_0 (the intercept) and b_1 for β_1 (the slope):

$$b_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad\qquad b_0 = \bar{Y} - b_1\overline{x}\;.$$

Introduction to Linear Models: Two Variables IV

Simple algebra yields

$$b_1 = r\frac{S_y}{S_x} = r\sqrt{\frac{S_{yy}}{S_{xx}}},$$

where r is the sample correlation coefficient

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{Y_i - \overline{Y}}{S_y}\right)\left(\frac{x_i - \overline{x}}{S_x}\right) = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(x_i - \overline{x})}{\sqrt{S_{xx}S_{yy}}}$$

where $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2, S_{yy} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2,$

and
$S_x = \sqrt{\frac{S_{xx}}{n-1}}, S_y = \sqrt{\frac{S_{yy}}{n-1}},$

Introduction to Linear Models: Two Variables V

Claim:

Like all least squares estimates for parameters in a linear model, the estimates are unbiased:

$E(b_0) = \beta_0$, and $E(b_1) = \beta_1$.

And,

$$Var(b_1) = \sigma^2 \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$Var(b_0) = Var(\bar{Y}) + \bar{x}^2 Var(b_1) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)$$

In the regression model, errors have distribution $\epsilon_i \sim (0, \sigma)$. So

$$b_1 \sim N\left(\beta_1, \frac{\sigma}{\sqrt{\sum(x_i - \bar{x})^2}}\right) \; b_0 \sim N\left(\beta_0, \sigma\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right)}\right)$$

Introduction to Linear Models: Two Variables VI

σ is unkown: calculate an unbiased estimate of the error variance: SSE/df

An unbiased estimate for σ^2 has the form

$$\widehat{\sigma}^2 \;=\; S^2 \;=\; \frac{\text{SSE}}{\text{df}} \;=\; \frac{\text{SSE}}{n-2} = \frac{\sum_{i=1}^{n} r_i^2}{n-2},$$

where each **residual** $r_i = Y_i - \hat{\mu}(x_i) = Y_i - (b_0 + b_1 x_i)$.

The degrees of freedom are $\text{df} = n - 2$ since n is the total sample size, and 2 is the number of parameters estimated in the model component for the mean (β_0, β_1).

Introduction to Linear Models: Two Variables VII

Compare models:

$Y_i = \mu + \epsilon_i$

$Y_i = \mu(x_i) + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i$

Do we need the more complex model?

Test or confidence interval for β_1:

$$H_o : \beta_1 = 0 \quad t\text{-statistic} = \frac{b_1 - 0}{SE(b_1)}$$

$$b_1 \pm t^* \, SE(b_1) \quad \text{where} \quad SE(b_1) = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

$t^* : P(T_{n-2} > t^*) = \alpha/2.$

Introduction to Linear Models: Two Variables VIII

CIs for the Mean Response

For a specific value of x, say x^*, the assumption is that y comes from a $N(\mu(x^*), \sigma)$ distribution, where

$$\mu(x^*) = \beta_0 + \beta_1 x^*$$

Plugging in our estimates of β_0 and β_1, $\mu(x^*)$ is estimated by $\hat{\mu}(x^*) = b_0 + b_1 x^*$, and a level $(1 - \alpha)$ confidence interval for the mean response $\mu(x^*)$ is given by

$$\hat{\mu}(x^*) \pm t^* \mathrm{SE}(\hat{\mu}(x^*))$$

where $t^* : P(T_{n-2} > t^*) = \alpha/2$.

$$\mathrm{SE}(\hat{\mu}(x^*)) = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Introduction to Linear Models: Two Variables IX

Prediction Interval for a Future Observation

Suppose we want to predict a specific observation value at $x = x^*$. At each x^*, $y \sim N(\mu(x^*), \sigma)$ We want to predict a y drawn from this distribution.

Our best guess is the estimated mean of the distribution,

$$\hat{Y} = \hat{\mu}(x^*) = b_0 + b_1 x^*$$

How accurate is this estimate?

The error here will be larger than the error for the mean response, $SE(\hat{\mu}(x^*))$, because there is error in estimating $\mu(x^*)$ as well as error in drawing a value from the normal distribution $N(\mu(x^*), \sigma)$.

Introduction to Linear Models: Two Variables X

A **level** $(1-\alpha)$ **prediction interval** for a future observation y corresponding to x^* is given by

$$\hat{Y} \pm t^* s_{\hat{Y}}$$

where $t^* : P(T_{n-2} > t^*) = \alpha/2$.

$$s_{\hat{Y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Analysis of variance I

Analysis of variance is the term for statistical analyses that break down the variation in data into separate pieces that correspond to different sources of variation. In the regression setting, the observed variation in the responses comes from two sources.

- As the explanatory variable x changes, it "pulls" the response with it along the regression line. This is the **variation along the line** or **regression sum of squares**:

$$\text{SSR} = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$$

- When x is held fixed, y still varies because not all individuals who share a common x have the same response y. This is the **variation about the line** or **error (residual) sum of squares**:

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Analysis of variance II

The ANOVA Equation

It turns out that SSE and SSR together account for *all* the variation in y (i.e. S_{yy}):

$$\underbrace{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{SSE}}$$

The degrees of freedom break down in a similar manner:

$$\underbrace{n-1}_{\text{SST}} = \underbrace{1}_{\text{SSR}} + \underbrace{n-2}_{\text{SSE}}$$

Dividing a sum of squares by its degrees of freedom gives a **mean square (MS)**.

Analysis of variance III

$$S^2 = \frac{\text{SSE}}{\text{df(Error)}} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$$

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\sum_i(\hat{Y}_i - \bar{Y})^2}{\text{SST}} = \frac{b_1^2 S_{xx}}{\text{SST}} = r^2$$

The fraction of the total variation in Y explained by the line

Multiple regression I

Suppose we have

- a single response variable y
- several predictor/explanatory variables $x_1, \ldots, x_p$

Data for multiple linear regression consist of the values of y and $x_1, \ldots, x_p$ for n individuals. We write the data in the form:

Individual	Predictors				Response
i	x_1	x_2	$\cdots$	x_p	Y
1	x_{11}	x_{12}	$\cdots$	x_{1p}	Y_1
2	x_{21}	x_{22}	$\cdots$	x_{2p}	Y_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{np}	Y_n

Multiple regression II

The multiple regression linear model posits the following relationship between y and $x_1, \ldots, x_p$:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

where

- $\epsilon \sim N(0, \sigma)$ is a random variable
- The ϵ_i's corresponding to observations $(Y_i; x_{i1}, x_{i2}, \cdots, x_{ip})$ on different individuals are independent of each other
- β_j is the change in y for each unit change in x_j *when holding all other predictors constant*

Multiple regression III

Estimating the Regression Parameters

The true population parameters $\beta_0, \beta_1, \ldots, \beta_p$ and σ are estimated from the data by the least squares method. That is, we minimize the *residual sum of squares*

$$\begin{aligned} \text{SSE} &= \sum_{i=1}^{n} (r_i)^2 \\ &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_{i1} - \cdots - b_p x_{ip})^2 \end{aligned}$$

Multiple regression IV

The estimator of σ^2 is

$$S^2 = \frac{\text{SSE}}{n-p-1} = \frac{\sum (r_i)^2}{n-p-1}$$

where $n-p-1$ is the number of degrees of freedom. (Why $n-p-1$?)

Multiple regression V

As with simple linear regression, we need to check that the model assumptions are met:

- The sample is a SRS from the population

  This can't be checked; this needs to be taken care of when the sample is drawn.

- There is a linear relationship in the population

  Checking this isn't as straightforward as with simple linear regression, but we should draw a plot of *residuals vs. fitted values* and check for any patterns.

- The standard deviation of the residuals is constant.

  Using the same plot as above, check for non-uniformity in the spread of residuals around the center line.

- The response varies Normally about the population regression line.

  Check with a *Normal quantile plot* of the residuals.

Multiple regression VI

Inference for Regression Coefficients

A 95% confidence interval for β_j is

$$b_j \pm t^* \mathrm{SE}(b_j)$$

where t^* is the number such that 95% of the area of the t_{n-p-1} distribution falls between $-t^*$ and t^*

To test the hypothesis

$$H_0 : \beta_j = 0 \qquad (\beta_i \text{ arbitrary for } i \neq j)$$

compute the t-statistic

$$T = \frac{b_j}{\mathrm{SE}(b_j)}$$

Multiple regression VII

- the p-value for this test statistic is computed from the t_{n-p-1} distribution
  $\rightarrow$ for $H_a : \beta_j > 0$, p-value is $P(t_{n-p-1} > T)$
  $\rightarrow$ for $H_a : \beta_j < 0$, p-value is $P(t_{n-p-1} < T)$
  $\rightarrow$ for $H_a : \beta_j \neq 0$, p-value is $2P(t_{n-p-1} > |T|)$
- if the regression model assumptions are true, testing $H_0 : \beta_j = 0$ corresponds to testing whether or not x_j is a significant predictor of y, *assuming all the other predictors are already in the model*.

Multiple regression VIII

ANOVA table for multiple regression

The basic ideas of the regression ANOVA table are the same in simple and multiple regression.

ANOVA expresses variation in the form of sums of squares. It breaks the total variation into two parts: SSR and SSE:

Source	SS	df
Regression (SSR)	$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$	p
Residual (SSE)	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$n - p - 1$
Total	$\sum_{i=1}^{n}(Y_i - \overline{Y})^2$	$n - 1$

$$\text{SST} = \text{SSR} + \text{SSE}$$

Multiple regression IX

The statistic

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = \frac{\text{SSR}}{\text{SST}} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

is the proportion of the variation of the response variable y that is explained by the explanatory variables $x_1, x_2, \cdots, x_p$. R^2 is called the **multiple correlation coefficient**.

Multiple regression X

The R^2 increases with every additional predictor. This is a mathematical fact. But some predictors may not be particularly useful in the regression.

Use Adjusted-R^2 :

$$R^2_{adj} = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)}$$

Adjusted R^2 does not necessarily increase with more predictors.

The adjusted R^2 compares the estimated sigmas - the numerator in the fraction is s. The denominator is fixed. So if s is smaller a model is better.

Comparing models - F test I

More generally, to test the hypothesis

H_0: q specific explanatory variables all have zero coefficients

H_a: at least one of the q has a nonzero coefficient

- Regress y on all predictor variables *except* the q variables of interest, and get the residual sum of squares SSE_{H_0}.
- Regress y on *all* predictor variables and get the residual sum of squares SSE_{H_a}.

Comparing models - F test II

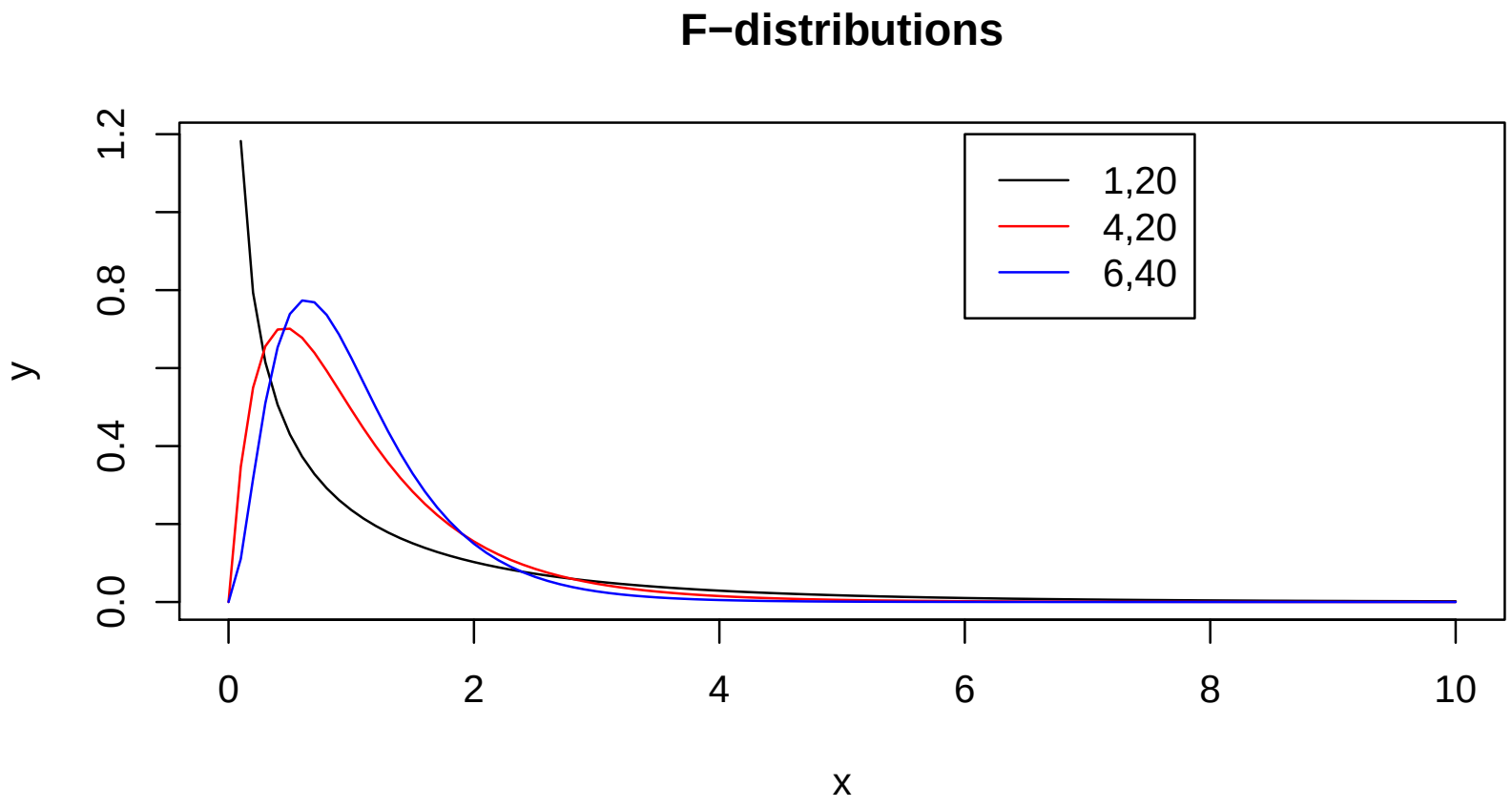
Calculate the F statistic:

$$
\begin{aligned}
F &= \frac{(\text{SSE}_{H_0} - \text{SSE}_{H_a})/(\text{df}_{H_0} - \text{df}_{H_a})}{\text{SSE}_{H_a}/\text{df}_{H_a}} \\
&= \frac{(\text{SSE}_{H_0} - \text{SSE}_{H_a})/q}{\text{SSE}_{H_a}/(n-p-1)} \\
&= \frac{(\text{SSE}_{H_0} - \text{SSE}_{H_a})/q}{S^2_{H_a}} \quad \rightarrow \text{decrease in SSE} \\
&= \frac{(\text{SSR}_{H_a} - \text{SSR}_{H_0})/q}{S^2_{H_a}} \quad \rightarrow \text{increase in SSR}
\end{aligned}
$$

Under H_0,

$$F \sim F_{q,n-p-1}$$

Comparing models - F test III



Binary predictors - dummy variables I

One of the predictors x could be binary. For example male/female.

Imagine two predicors x_1, x_2 - one continuous and one binary.

The linear model is:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}.$$

We can write this:

$$Y_i = \begin{cases} \beta_0 + \beta_1 x_{i1} & \text{if } x_{i2} = 0 \\ \beta_0 + \beta_2 + \beta_1 x_{i1} & \text{if } x_{i2} = 1 \end{cases}$$

This produces two parallel lines one for each level of x_2. β_2 is the increment in intercept for $x_2 = 1$.

Binary predictors - dummy variables II

If we add an interaction between the continuous variable and the binary variable: $x_{i1} \cdot x_{i2}$, we get a model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3},$$

which can be written in terms of cases as follows:

$$Y_i = \begin{cases} \beta_0 + \beta_1 x_{i1} & \text{if } x_{i2} = 0 \\ \beta_0 + \beta_2 + (\beta_1 + \beta_3) x_{i1} & \text{if } x_{i2} = 1 \end{cases}$$

So β_2 gives the increment in the intercept for the category $x_2 = 1$ and β_3 the increment in the slope for $x_2 = 1$.