Multiple Linear Regression.

A. Skripnikov[1]

[1]New College of Florida

IDC 5205

Multiple Linear Regression: Two Variables.

Multiple linear regression deals with **multiple** explanatory variables.

Example. Now, let's include both $X_1 = TV$ and $X_2 = radio$ to predict $Y = sales$. Then the multiple linear regression modeling equation is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon, \ \epsilon \sim_{i.i.d.} N(0, \sigma^2),$$

or (with "i" indexation)

Replacing "Y" and "X" with the actual names of variables:

or (with "i" indexation):

Multiple Linear Regression: Two Variables.

Corresponding multiple linear regression equation for **predictions** is:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Task: Find $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ such that

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \hat{\beta}_2 X_{2,i} \text{ (the } \textbf{predicted} \text{ value)}$$

is as close as possible to

$$Y_i \text{ (the } \textbf{true} \text{ value)}, \ i = 1, \dots, n$$

We need to minimize magnitude of residuals $e_i = Y_i - \hat{Y}_i, \ i = 1, \dots, n$

That's done via **least squares** yet again: $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ result from

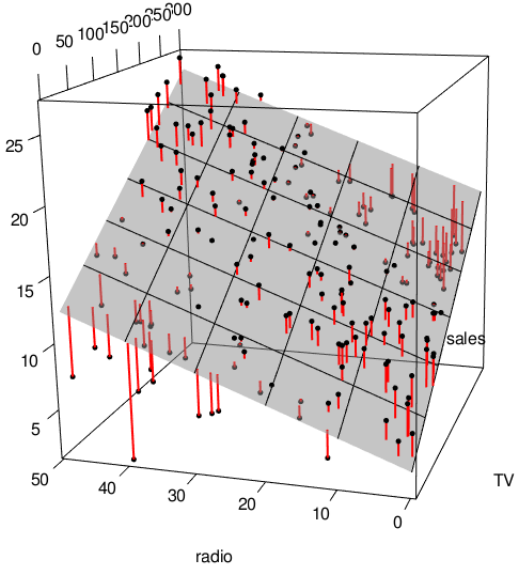
$$\underset{\beta_0, \beta_1, \beta_2}{min} \ RSS =$$

Geometry of Least Squares for 2-Predictor Regression.

Geometrically, it amounts to finding a **plane**

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

that's closest to the data points (vertical lines are residuals $e_i = \hat{Y}_i - Y_i$):



Slopes in Multiple Linear Regression: Partial Effects.

Development of least-squares linear regression estimates is similar for simple and multiple regression, but their **interpretation** is **different**:

- $\hat{\beta}_1$ in **multiple** regression is a **partial** coefficient: it represents the "effect" on Y of a one-unit increment in X_1, **holding** X_2 **constant**. Same goes for $\hat{\beta}_2$, but **holding** X_1 **constant**.

- The $\hat{\beta}_1$ in **simple** regression represents the **marginal** relationship between Y and X_1, completely ignoring X_2.

Advertising example.

Task (see *R* code). Fit the following multiple linear regression model for *Advertising* data:

$$sales \sim TV + radio$$

and

1. Write down the fitted model equation.

2. Interpret the coefficients.

   For *TV* predictor, $\hat{\beta}_1 = 0.045$: Per $1,000$\$ increase in *TV* budget, **holding** *radio* **budget constant**, ...

Interpreting Effects of Predictors in MLR.

Example (cont'd). Fitted MLR equation is:

$$\widehat{sales} = 2.921 + 0.046 \times TV + 0.188 \times radio$$

Generic version for **slope interpretation in MLR**:

Per 1-unit increase in X, **holding all other predictors constant**, Y will increase ($\hat{\beta} > 0$)/decrease ($\hat{\beta} < 0$) by $|\hat{\beta}|$ units, **on average**.

Task. Interpret effects of each advertising media on sales.

- For *radio*, $\hat{\beta}_2 = 0.189$:

Interpreting the **intercept** in MLR.

Example (cont'd). Fitted MLR equation is:

$$\widehat{sales} = 2.921 + 0.046 \times TV + 0.188 \times radio$$

- For *intercept*, $\hat{\beta}_0 = 2.921$:

Multiple Linear Regression (MLR): General p Predictors.

Full modeling equation for multiple linear regression with p **predictors**:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \ \epsilon \sim_{i.i.d.} N(0, \sigma^2),$$

or, with proper "i" indexation,

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_p X_{p,i} + \epsilon_i, \ \epsilon_i \sim_{i.i.d.} N(0, \sigma^2)$$

Model assumptions on conditional response distribution $(Y \mid X_1, \dots, X_p)$ are **analogous** to those of **simple linear regression**.

Q #1: What were those modeling assumptions?

Multiple Linear Regression (MLR): General p Predictors.

The fitted p-dimensional **hyperplane** will have the following equation:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

Although impossible to visualize the data with a p-dimensional plane (p$\geq$ 3), estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are still found via **least squares**.

Q #2: Assumptions from **Q #1** lead to **what properties** of least squares estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ for population parameters $\beta_0, \beta_1, \ldots, \beta_p$?

Statistical Inference for Multiple Linear Regression.

Just like for simple linear regression,

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{V(\hat{\beta}_j)}} \sim N(0,1), \qquad j = 0, 1, 2, \ldots, p$$

where $V(\hat{\beta}_j)$ contains the unknown σ, which we substitute for $\hat{\sigma}$:

$$\hat{\sigma} = RSE = \sqrt{\frac{RSS}{n-(p+1)}} = \sqrt{\frac{\sum_i (Y_i - \hat{Y}_i)^2}{n-(p+1)}} \tag{1}$$

That leads to us using t-distribution:

$$\frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \sim t_{n-(p+1)}, \qquad j = 0, 1, 2, \ldots, p \tag{2}$$

Q: Why "$n-(p+1)$" in (1) and (2)?

Statistical Inference in MLR: Confidence Intervals.

Fact (2) leads to the $(1-\alpha)\%$ **confidence interval** formula for β_j:

$$(\hat{\beta}_j - t_{[n-(p+1),1-\alpha/2]} SE(\hat{\beta}_j),\ \hat{\beta}_j + t_{[n-(p+1),1-\alpha/2]} SE(\hat{\beta}_j))$$

Task (See *R* code). Fit multiple linear regression of

$$sales \sim TV + radio$$

Next, for $\boldsymbol{\beta_1}, \boldsymbol{\beta_2}$, proceed to

- Calculate and **interpret** the 95% and 90% confidence intervals.

Statistical Inference in MLR: Confidence Intervals.

Task (cont'd).

- Calculate and **interpret** the 95% and 90% confidence intervals.

Statistical Inference in MLR: Hypothesis Testing.

Task (cont'd). In multiple linear regression of

$$sales \sim TV + radio$$

for $\boldsymbol{\beta_1}, \boldsymbol{\beta_2}$, proceed to

- Interpret the *summary*() results of respective hypotheses tests.

Multiple Linear Regression: Advertising example.

Task. For the following multiple linear regression model:

$$sales \sim TV + radio + newspaper,$$

proceed to:

- Write down the general modeling equation.
- Fit the model in R, write the **fitted** model equation.
- Interpret the coefficients.

  For TV, $\hat{\beta}_1 = 0.046$:

Task (cont'd).

- Interpret the coefficients (cont'd).

Advertisement example.

Example. Having fitted MLR for *Advertisement* data, we got

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Questions:

- **Practically**, what does **standard error** describe for each coefficient?
- What H_0 hypotheses do the t-statistic and p-value columns refer to?

Advertisement example.

Example. Having fitted MLR for *Advertisement* data, we got

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	2.939	0.3119	9.42	< 0.0001
`TV`	0.046	0.0014	32.81	< 0.0001
`radio`	0.189	0.0086	21.89	< 0.0001
`newspaper`	-0.001	0.0059	-0.18	0.8599

Questions (cont'd):

- What values of t-statistic are expected of H_0 were true?
- What do the p-values tell us about H_0?

Newspaper advertising in SLR & MLR.

As a result of fitting **multiple** linear regression, we see that *newspaper* doesn't have an effect on sales, **BUT...**

Meanwhile, in **simple** linear regression:

$$sales = \beta_0 + \beta_1 \times newspaper + \epsilon,$$

we actually witness a **significant effect** for *newspaper*:

```
> lm.news <- lm(sales ~ newspaper, data=Advertising)
> summary(lm.news)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.35141    0.62142   19.88  < 2e-16 ***
newspaper    0.05469    0.01658    3.30  0.00115 **
```

Q: **Why?**

A: **See** *R* **code** and the following slides...

Newspaper advertising in SLR & MLR.

Example (cont'd): Full correlation matrix for *Advertisement* data set

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

What do we see?

$cor(newspaper, radio) =$

It means: in simple linear regression of $sales \sim newspaper$, a $1,000$\$ increase in newspaper ads is also accompanied by ...

Example (cont'd):

Q: Could we make that mistake in MLR? Why? **See** R **code**.

Absurd example: Shark Attacks & Ice Cream sales.

Example:

- **Simple** linear regression of $Y = \{\#\text{ shark attacks}\}$ on $X_1 = \{\text{ice cream sales}\}$ might yield a **strong positive relationship**.
- Extra variable correlated with both is $X_2 = \{\textbf{\# of beach visitors}\}$:
  - more people at the beach $\Longrightarrow$ more shark attacks & ice cream sold

Question: What would likely happen in a **multiple** linear regression of shark attacks (Y) on both ice cream sales (X_1) & $\#$ of beach-goers (X_2)?

Question: How do we match up the variables from *Advertisement* example with those in the *Shark* example in terms of their roles?

Multiple Linear Regression: Important Questions.

Critical questions when performing MLR:

1. (**Model significance**) Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
2. (**Variable selection**) Do all the predictors help to explain Y, or is only a subset of the predictors useful?
3. (**Quality of Fit**) How well does the model fit the data?
4. (**Prediction**) Given a set of predictor values, what response value should we predict, and how confident are we in our prediction?

Model significance: F-statistic.

To determine if at least one predictor has a strong relationship with the response (to **test for model significance**):

1. Hypotheses are

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad vs \quad H_a : \{\text{at least one } \beta_j \neq 0\}$$

2. Test statistic used is ***F*-statistic (FS)**:

$$FS = \frac{RegSS/p}{RSS/(n-(p+1))}$$

where

- RSS (Residual Sum of Squares) $= \sum_i (Y_i - \hat{Y}_i)^2$ - {initial variance in response Y that's left **un**explained after regression},
- $RegSS$(Regression Sum of Squares) $= TSS - RSS =$ {initial variance in response Y that's **explained** by regression}.

Model significance: Analysis of Variance (ANOVA).

This breakdown of the **initial variance in** Y (TSS) into

- Variance left **un**explained by regression (RSS),
- Variance **explained** by regression ($RegSS$),

$$TSS = RSS + RegSS,$$

is known as **ANalysis Of VAriance (ANOVA)** for **regression**.

In many traditional statistics texts, it's accompanied by an ANOVA table:

Source	*Sum of Squares*	*df*	*Mean Square*	*F*
Regression	RegSS	p	$\frac{\text{RegSS}}{p}$	$\frac{\text{RegMS}}{\text{RMS}}$
Residuals	RSS	$n-p-1$	$\frac{\text{RSS}}{n-p-1}$	
Total	TSS	$n-1$		

F-Statistic Breakdown.

$$FS = \frac{RegSS/p}{RSS/(n-(p+1))}$$

Given that

- $RegSS$ = initial variance in response Y that's ...

- RSS = initial variance in response Y that's ...

Question: What values of F-statistic indicate a **good model** - high or low? Why?

F-test.

To formalize **how high** a value of F-statistic is **evidence enough** to claim **model significance** ($\Leftrightarrow$ reject $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$), we need **sampling distribution of F-statistic** under H_0.

3. Under $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$, we would expect:

$$(F \mid H_0) \quad \sim \quad F_{p,n-(p+1)},$$

   where $\boldsymbol{F_{p,n-(p+1)}}$ is $\boldsymbol{F}$-distribution with
   - p **numerator** degrees of freedom, and
   - $n - (p+1)$ **denominator** degrees of freedom

   See `https://istats.shinyapps.io/FDist/` **for demo.**

4. p-**value**$=$ {how likely is $F = FS$ value (or more extreme) under H_0}:

   p-value $= P(F \geq FS \mid H_0) \equiv P(F_{p,n-(p+1)} \geq FS)$

   Question: Why does "more extreme" mean **higher** values of F-stat?

Sampling Distribution of F-statistic (**FOR CURIOUS**).

Definition. Random variable F (for F-statistic) has F-distribution with p numerator and $n-(p+1)$ denominator degrees of freedom (AKA $FS \sim F_{p,n-(p+1)}$)

$$F = \frac{X_p^2/p}{X_{n-(p+1)}^2/(n-(p+1))},$$

where $X_p^2 \sim \chi_p^2,\ X_{n-(p+1)}^2 \sim \chi_{n-(p+1)}^2,\ \ X_p^2\ \&\ X_{n-(p+1)}^2$ are **independent**.

Definition. Random variable X_p^2 has χ^2 distribution with p degrees of freedom (AKA $\sim \chi_p^2$) if

$$\chi_p^2 = Z_1^2 + Z_2^2 + \cdots + Z_p^2,$$

where $Z_i \sim_{i.i.d} N(0,1),\ \ i = 1, \ldots, p$.

Sampling Distribution of F-statistic (**FOR CURIOUS**).

In our case, from some **deep linear models theory**, the following result on sampling distributions of $RegSS$ and RSS **under** H_0 is available:

$$RegSS = \sum_i (\hat{Y}_i - \bar{\mathbf{Y}})^2 \mid H_0 \sim \sigma^2 \chi^2_p$$

$$RSS \quad = \sum_i (Y_i - \hat{Y}_i)^2 \mid H_0 \sim \sigma^2 \chi^2_{n-(p+1)},$$

leading to

$$F = \frac{RegSS/p}{RSS/(n-(p+1))} \equiv$$

$$\frac{RegSS/p}{RSS/(n-(p+1))} \mid H_0 \quad \sim \quad \frac{\cancel{\sigma^2} \chi^2_p / p}{\cancel{\sigma^2} \chi^2_{n-(p+1)} / (n-(p+1))} \equiv F_{p,n-(p+1)}$$

P**-value** - how likely is this value of FS under H_0 - is calculated as

$$\text{p-value} = P(F_{p,n-(p+1)} \geq FS)$$

Steps of F-test.

Task. Proceed to lay out the steps of F-test in general case (hypotheses, test statistic & its sampling distribution, p-value calculation, conclusion).

F-Test for *Advertisement*.

Task (See R **code as well)**. Proceed to lay out the steps of F-test for *Advertisement* example of

$$sales \sim TV + radio + newspaper$$

Why need F-statistic?

Question: Why not just look at **single variable** t-test results and p-values (if at least one is significant - reject H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$)?

Answer: In many cases - sure, but **generally WRONG.**

Example. Say, for

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{100} X_{100} + \epsilon,$$

we know that H_0: $\beta_1 = \beta_2 = \cdots = \beta_{100} = 0$ is indeed **true**.

Then, at significance level $\alpha = 0.05$, we can expect 5% of p-values ending up < 0.05 simply due to **Type I error** allowance:

$$\alpha = P(\text{Reject } H_0\text{: } \beta_j = 0 \mid H_0 \text{ is true}) = P(\text{Type I error}), \; j = 1, 2, \ldots, 100$$

Meanwhile, ***F*-test doesn't suffer from this issue**, by looking at the effect of the **model as a whole**.

See the homework problem on "Why need F-statistic?".

Why need F-statistic? Rule of thumb.

1. If F-test shows that the model is insignificant: **drop the model**.

   Issues might be:

   - predictors are really useless,
   - relationship is not linear, hence some data transformations are needed (see next deck of slides).

2. If F-test shows that the model is significant: proceed to interpret the predictors with significant individual t-tests.

F-test: Reverse situation.

NOTE: Reverse situation is also possible in multiple linear regression.

Example. Data on 25 patients with cystic fibrosis yields:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 176.0582   225.8912   0.779    0.448
age          -2.5420     4.8017  -0.529    0.604
sex          -3.7368    15.4598  -0.242    0.812
height       -0.4463     0.9034  -0.494    0.628
weight        2.9928     2.0080   1.490    0.157
bmp          -1.7449     1.1552  -1.510    0.152
fev1          1.0807     1.0809   1.000    0.333
rv            0.1970     0.1962   1.004    0.331
frc          -0.3084     0.4924  -0.626    0.540
tlc           0.1886     0.4997   0.377    0.711
...
F-statistic: 2.929 on 9 and 15 DF,  p-value: 0.03195
```

Task. Comment on what you observe.

Issue of **Collinearity**.

Example (cont'd). The overall model significance tells us that **there is at least one important variable** in the set of p predictors,

BUT

We can't see its' **single** t**-test** significance yet due to the

- **Issue** of **collinearity**:

  some important predictors are strongly correlated with each other

$$\Downarrow$$

makes it difficult to separate their effects on the response.

Solution: **drop variables** that cause collinearity.

Collinearity.

In linear regression, **collinearity** refers to two or more predictors being closely **linearly** related to each other.

Example. In *Credit* data set, using customers' credit *Balance* as response, and customer's *Age*, *Limit* and *Rating* as predictors, the **predictor correlation matrix** looks as follows:

	Age	Limit	Rating
Age	1.000	0.101	0.103
Limit	0.101	1.000	0.997
Rating	0.103	0.997	1.000

Question: Which predictors are collinear? Why?

Collinearity.

Example (cont'd). Let's fit the following two models:

1. $Balance = \beta_0 + \beta_{Age} \times Age + \beta_{Limit} \times Limit + \epsilon$,
2. $Balance = \beta_0 + \beta_{Rating} \times Rating + \beta_{Limit} \times Limit + \epsilon$

Question: Which model has collinearity? Why?

Example (cont'd). Let's check the impact collinearity may have on inference, by looking at the hypothesis test results for these two models:

		Coefficient	Std. error	t-statistic	p-value
Model 1	`Intercept`	−173.411	43.828	−3.957	< 0.0001
	`age`	−2.292	0.672	−3.407	0.0007
	`limit`	0.173	0.005	34.496	< 0.0001
Model 2	`Intercept`	−377.537	45.254	−8.343	< 0.0001
	`rating`	2.202	0.952	2.312	0.0213
	`limit`	0.025	0.064	0.384	0.7012

Q: What happens to $SE(\hat{\beta}_{limit})$ and corresponding p-value once collinearity is introduced? **See** R **code as well**.

Collinearity.

The uncertainty introduced by collinearity is reflected in increased standard error of respective coefficient estimates.

Q: Why is it **intuitive** that collinearity makes estimating β_j coefficients tougher? **Hint**: Recall $\hat{\beta}_j$ interpretation in multiple linear regression.

Collinearity.

Presume we have the following "fitted" model:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2,$$

where X_1, X_2 are **perfectly correlated** ($cor(X_1, X_2) = 1$).

Q: Can we estimate β_1, β_2 here? Why? **Hint**: Recall their interpretations.

Collinearity: Remedial measures.

There are a couple approaches to deal with collinearity:

Approach #1: Correlation matrix of predictors.

1. Calculate full correlation matrix of all predictors.
2. From a group of variables that are highly correlated (e.g. $|cor| > 0.90$, or 0.80) with each other, proceed to only retain one of them in the model, while dropping the others.

See R **code example.**

Multi-Collinearity.

Issue with Approach #1: it is possible for collinearity to exist between **three or more** variables even if there are **no high pairwise correlations**. That is known as **multi-collinearity**.

Example. For *Advertising* data,

1. Let $Total = radio + newspaper$,
2. Consider regression $sales \sim radio + newspaper + total$.

Task. Work through this example (**see** R **code as well**).

Multi-Collinearity: Variance Inflation Factor (VIF).

Question: How to automatically detect that "sneaky" multi-collinearity?

Answer: **Variance Inflation Factor (VIF)**.

Variance Inflation Factor (VIF) for predictor X_j is calculated as

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

where $R^2_{X_j|X_{-j}}$ is the R^2 from regressing X_j onto all **other predictors**:

$$X_j \sim X_1 + \cdots + X_{j-1} + X_{j+1} + \cdots + X_p$$

If we get

- $R^2_{X_j|X_{-j}} \approx 1 \implies$
- $R^2_{X_j|X_{-j}} \approx 0 \implies$

Multi-Collinearity: Variance Inflation Factor (VIF).

Approach #2: Variance Inflation Factor (VIF). Proceed to

1. Calculate VIF for each predictor in the model.
2. In case there are VIF values ≥ 5 (**rule of thumb**) - drop the **one** predictor corresponding to the **largest VIF value**.
3. Repeat steps 1 & 2 for reduced model, until **all VIF values are** < 5.

See *R* code for examples.

Variable (Model) Selection.

Even after collinearity gets taken care of, one should still further consider:

Variable (or model) selection - task of retaining in the model only the **most essential variables**.

Variable selection methods differ by the

1. direction of the procedure:
   - **Backward**: starts with *full* model, *drops* variables one at a time,
   - **Forward**:
   - **Mixed**: combines **forward** and **backward**.
2. selection criteria used:
   - Akaike Information Criteria (AIC),
   - Bayesian Information Criteria (BIC),
   - Mallow's C_p

Here, we'll just provide the default approach of *step*() function in *R*:

backward selection via **AIC** criteria

Backward Selection via **AIC**.

1. Start with a full model $\{\beta_1, \beta_2, \ldots, \beta_p\}$.
2. Drop one variable (and its coefficient β_j) at a time, aiming to **minimize** an information criterion AIC:

$$AIC(\beta_1, ..., \beta_p) \approx \{\text{Model Fitting Error}\} + \{\text{Model Complexity}\} \approx$$
$$\approx RSS(\beta_0, \beta_1, ..., \beta_p) + (p+1)$$

   which accounts for

   - model **fit quality** -

   - model **complexity** -

   Note: Dropping a variable from the model:
   - always makes the fit quality worse (larger RSS), **BUT**
   - it also decreases the complexity,

   hence AIC balances those out to **select** the best subset of variables.
3. Stop once you can't improve AIC by dropping any of the remaining variables. **See** *step*() **function output for illustration**.

Backward Selection via **AIC**.

Example. Using backwards AIC selection on *cystfibr* data:

```
> lm.obj <- lm(pemax ~ ., data=cystfibr)
> step(lm.obj)
Start:  AIC=169.11
pemax ~ age + sex + height + weight + bmp + fev1 + rv + frc + tlc
         Df Sum of Sq      RSS    AIC
- sex     1     37.90   9769.2 167.20
- tlc     1     92.40   9823.7 167.34
...
<none>                  9731.2 169.11
...


Step:  AIC=167.2
pemax ~ age + height + weight + bmp + fev1 + rv + frc + tlc

         Df Sum of Sq      RSS    AIC
- tlc     1    115.94   9885.1 165.50
- height  1    131.21   9900.4 165.54
...
```

Backward Selection via **AIC**.

Example (ct'd). After *several more steps*:

```
...
Step:  AIC=160.66
pemax ~ weight + bmp + fev1 + rv

         Df Sum of Sq   RSS    AIC
<none>                10355 160.66
- rv      1    1183.6 11538 161.36
- bmp     1    3072.6 13427 165.15
- fev1    1    3717.1 14072 166.33
- weight  1   10930.2 21285 176.67
```

Qs:

- What's the final model selected?

- Why did the *step*() function stop at that particular model?

Quality of Fit: RSE and R^2 for Multiple Linear Regression.

To measure the **quality of fit** for our multiple linear regression model, we use the following metrics:

- Residual Standard Error (RSE):

$$RSE = \sqrt{\frac{SSE}{n-(p+1)}} = \sqrt{\frac{\sum_i e_i^2}{n-(p+1)}}$$

- R^2 (coefficient of determination) calculated in the same manner:

$$R^2 = \frac{TSS - RSS}{TSS} = \frac{\sum_{i=1}^{n}(Y_i - \bar{\mathbf{Y}})^2 - \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{\mathbf{Y}})^2}$$

Quality of Fit: RSE and R^2 for Multiple Linear Regression.

Task (see *R* code). For advertisement data, when fitting the full model

$$sales \sim TV + radio + newspaper$$

proceed to interpret

- $RSE = ...$

- $R^2 = ...$

Prediction for MLR: Prediction and Confidence Bands.

Last, but not least, the question of **prediction** for MLR:

Example (See *R* code). Presume we spend $50,000\$$ on TV, $20,000\$$ on radio and $5,000\$$ on newspaper ads. **Interpret the following statements**:

- *sales* prediction is .... $\Longrightarrow$
- *sales* 95% confidence bands are .... $\Longrightarrow$
- *sales* 95% prediction bands are .... $\Longrightarrow$

Dealing with Categorical Variables.

So far we've mostly dealt with **quantitative** predictors: advertisement budget, age, height, weight,...

Q: How does one incorporate **categorical** predictors?

Example. *Carseats* data set deals with car seat sales in different stores based on a variety of area and store characteristics. Besides

- **quantitative** variables (e.g. income, competitor price, population ...),

it also utilizes

- **categorical** variables

such as (**see** *R* **code**)

- *Urban* (Yes/No),
- ...
- ...

Categorical Predictors with Two Levels.

Example (ct'd). Say, we wish to investigate differences in car seat sales for **urban** and **non-urban** areas, ignoring other variables for the moment.

We create a **dummy variable** for *Urban* status of the area:

$$D_{Urb,i} = \begin{cases} 1, & \text{if } i^{th} \text{ store is in urban area,} \\ 0, & \text{otherwise,} \end{cases}$$

and use it as a **predictor** in SLR, resulting into

$$Sales_i = \beta_0 + \beta_1 D_{Urb,i} + \epsilon_i =$$

Using **least squares**, one would obtain (**see** *R* **code to confirm**)

- $\hat{\beta}_0$ - average car seat sales for stores in **non-urban** areas,
- - average car seat sales for stores in **urban** areas.

Categorical Predictors with Two Levels.

Example (ct'd). For our "single dummy variable" model,

$$Sales_i = \beta_0 + \beta_1 D_{Urb,i} + \epsilon_i,$$

how to specify a hypothesis test for the following question

"Is there a significant difference in car seat sales for stores in urban and non-urban areas?"

Categorical Predictors with Two Levels.

Example (cont'd). Below is the coefficients table:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.56356    0.26028  29.060   <2e-16 ***
UrbanYes    -0.09537    0.30998  -0.308    0.759
```

Task. Proceed to interpret:

- The effect of urban area on car seat sales.
- Std. error for that effect.
- Hypothesis test results for that effect

Example (cont'd). Below is the coefficients table:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.56356    0.26028  29.060   <2e-16 ***
UrbanYes    -0.09537    0.30998  -0.308    0.759
```

Task. Proceed to interpret:

- *Intercept* coefficient
- Std. error for that effect.
- (if relevant) hypothesis test results for *Intercept*

Categorical Predictors with > 2 Levels.

Q: What if a variable has > 2 **categories**?

A: Create <u>**more**</u> **dummy variables**.

Example. To model the shelve location variable with $K = 3$ levels (Good, Medium, Bad) in *Carseats* data set, we use $K - 1 = 2$ **dummy variables:**

$$D_{GoodLoc,i} = \begin{cases} 1, & \text{shelve location in } i^{th} \text{ store is good} \\ 0, & \text{otherwise} \end{cases},$$

$$D_{MedLoc,i} = \begin{cases} 1, & \text{shelve location in } i^{th} \text{ store is medium} \\ 0, & \text{otherwise} \end{cases}.$$

Note: *Bad* location is our **baseline** (or **reference**) category here.

Categorical Predictors with > 2 Levels.

Example (ct'd). With $D_{GoodLoc}$ and D_{MedLoc} from previous slide, we get:

$$Sales_i = \beta_0 + \beta_1 D_{GoodLoc,i} + \beta_2 D_{MedLoc,i} + \epsilon_i =$$

Categorical Predictors with > 2 Levels.

Task. For our model of

$$Sales_i = \beta_0 + \beta_1 D_{GoodLoc,i} + \beta_2 D_{MedLoc,i} + \epsilon_i,$$

specify hypotheses in order to answer the following questions:

1. **"Is there a significant difference in car seat sales between good and bad shelve locations?"**

2. **"Does shelve location affect car seat sales?"**

Task (ct'd). Fit the $Sales \sim ShelveLoc$ regression model,

- Write down the fitted equation.

- Interpret all $\hat{\beta}$ coefficients. **Should we trust all of these interpretations (from statistical significance standpoint)**?

Categorical Predictors with > 2 Levels.

Task (cont'd). Fit the *Sales* $\sim$ *ShelveLoc* regression model,

- Interpret all $\hat{\beta}$ coefficients. **Should we trust all of these interpretations (from statistical significance standpoint)**?

Categorical Predictors in MLR.

Example. Lastly, let's run MLR with a **mixture** of quantitative & categorical predictors:

$$Sales \sim Advertising + ShelveLoc$$

Task.

- Write down full modeling equation.

- Write down the fitted equation.

Categorical Predictors in MLR.

Task (cont'd).

- Interpret all β's (including β_0). **Should we trust all of these interpretations (from statistical significance standpoint)**?

Categorical Predictors in MLR.

Task (cont'd).

Testing for Significance of a Categorical Predictor.

Q: How could we test for significance of a categorical predictor, when being used in multiple linear regression with other predictors, if it

- has $K = 2$ levels? **See** R **code**.

- has $K > 2$ levels?

  Option #1 (wrong one):

  Q: Why is that approach wrong?

  Option #2 (correct one): Use **incremental** F**-test** for significance of a **subset** of predictors.

Testing for Significance of a Categorical Predictor.

Example. Full model eq. for $Sales \sim Advertising + ShelveLoc$ regression

$$Sales_i = \beta_0 + \beta_1 Advert_i + \beta_2 D_{GoodLoc,i} + \beta_3 D_{MedLoc,i} + \epsilon_i, \; \epsilon_i \sim_{i.i.d.} N(0, \sigma^2),$$

Task: Formulate the hypotheses that will address the significance of the **entire** *ShelveLoc* categorical variable.

Hence, technically, we want to test for significance of **subset** of predictors (in that case, $D_{GoodLoc}$ and D_{MedLoc}).

Q: How?

A: **Incremental** F-test.

Incremental F-test: Significance for Subset of Predictors.

Incremental F-test for hypotheses

$$H_0\colon\ \beta_2 = \beta_3 = 0, \quad vs \quad H_a\colon\ \{\text{at least one of } \beta_2, \beta_3 \neq 0\}$$

in regression like this

$$Sales_i = \beta_0 + \beta_1 Advert_i + \beta_2 D_{GoodLoc,i} + \beta_3 D_{MedLoc,i} + \epsilon_i,\ \epsilon_i \sim_{i.i.d.} N(0, \sigma^2),$$

is based on a comparison of **RegSS** (**Q**: What's that?) for

- **Full** model: includes **all** slope coefficients ($RegSS_{Full}$)

- **Null** model, with β_2, β_3 **excluded** ($RegSS_{Null}$). It is **nested** within the full model (hence the name "incremental" F-test).

Incremental F-test: Significance for Subset of Predictors.

The incremental F-test statistic is:

$$FS = \frac{(RegSS_{Full} - RegSS_{Null})/q}{RSS_{Full}/(n-(p+1))}$$

where we have

- q - number of β-coefficients tested in H_0 (e.g. if H_0: $\beta_2 = \beta_3 = 0$, then $q = 2$)
- numerator is the **incremental** sum of squares, capturing the **increase** in response variance **explained** when going from "Null" $\Rightarrow$ "Full" model (as in - when adding that subset of predictors),
- denominator - unbiased estimate of error variance (details left out)

Q: Intuitively, what values of FS serve as evidence to reject H_0?

Incremental F-test: Significance for Subset of Predictors.

Q: How high an F-statistic value is evidence enough to reject H_0?

A: Need **sampling distribution** of F-stat vals **expected under** H_0.

Under H_0 being true, we have

$$F \mid H_0 \ \sim \ F_{q,n-(p+1)}$$

The **exact** p-**values** - quantifying the evidence of how likely it was to see such value of $F = FS$ (or more extreme) - are calculated as

$$\text{p-value} = P(F_{q,n-(p+1)} \geq FS)$$

Steps of **incremental** F-test.

Task. Proceed to lay out the steps of incremental F-test in general case (hypotheses, test stat. & its sampling distr., p-value calc., conclusion).

Incremental F-test: Significance for Subset of Predictors.

Example (ct'd). For our $Sales \sim Advertising + ShelveLoc$ regression:

$$Sales_i = \beta_0 + \beta_1 Advert_i + \beta_2 D_{GoodLoc,i} + \beta_3 D_{MedLoc,i} + \epsilon_i, \ \epsilon_i \sim_{i.i.d.} N(0, \sigma^2),$$

lay out the steps of incremental F-test for significance of $ShelveLoc$.

See R code for example.

Incremental F-test: Significance for Subset of Predictors.

Example (cont'd). Be able to interpret all entires in the summary below:

```
> anova(lm.null.obj, lm.obj)
...
Model 1: Sales ~ Advertising
Model 2: Sales ~ Advertising + ShelveLoc
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    398 2951.1
2    396 1994.4  2    956.74 94.984 < 2.2e-16 ***
```