Linear Regression

Yang Wang

College of Charleston

Linear regression model

Linear regression model assumes that the dependence of y on $x_1, x_2, ..., x_p$ is linear.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

- y: response, dependent variable
- $\boldsymbol{x} = (x_1, ..., x_p)$: regressors, independent variables, predictors
- $\boldsymbol{\beta} = (\beta_0(\mathit{intercept}), \beta_1, ..., \beta_p(\mathit{slops}))$: regression coefficients, parameters
- ϵ: error term, noise

Although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.

Assumptions of linear regression model

- The choice of predictors and their form is correct (linearity)
- The records are independent of each other
- The variability in the outcome values for a given set of predictors is the same regardless of the values of the predictors (equal variance or homoscadasticity)
- The noise ϵ follows a normal distribution with mean 0 and variance σ^2 (this assumption is needed for estimating regression coefficients)
- As a result of above y also follows a normal distribution with mean $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ and variance σ^2.

Linear regression model

For a n p-variate data points $\{y_i, x_{i1}, ..., x_{ip}\}$ for $i = 1, ..., n$

- $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$
- y_i: observed response for observation i
- x_{ij}: value of j^{th} regressor for observation i
- ϵ_i is the noise or unexplained part with mean 0 and variance σ^2
- $E(y_i|\boldsymbol{x} = (x_{i1}, ..., x_{ip})) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$
- A model linear in parameters as well as in predictors.

Parameters Estimation

- The estimates of $\beta_0, \beta_1, \ldots \beta_p$ are obtained using the method of ordinary least squares (OLS).
  - Consider the deviation of y_i from it's expected value: $e_i = y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi})$ (ith residuals)
  - Sum of squared errors (SSE) as SSE $= e_1^2 + ... + e_n^2$
  - Minimize the sum of the squared deviations SSE $= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_p x_{pi})^2$ for the given sample observations
    - recall: the derivative $f'(x)$ is 0 at point x at which $f(x)$ is a maximum or minimum
    - solve SSE$' = 0$ to find minima $\boldsymbol{\beta}$ that is the estimated parameters $\hat{\boldsymbol{\beta}}$
- The estimate of σ^2 is $MSE = \frac{SSE}{n-(p+1)}$

Parameters Estimation

- Matrix representation of MLR model: $\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{p1} \\ 1 & x_{12} & x_{22} & \dots & x_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{pn} \end{bmatrix},$$

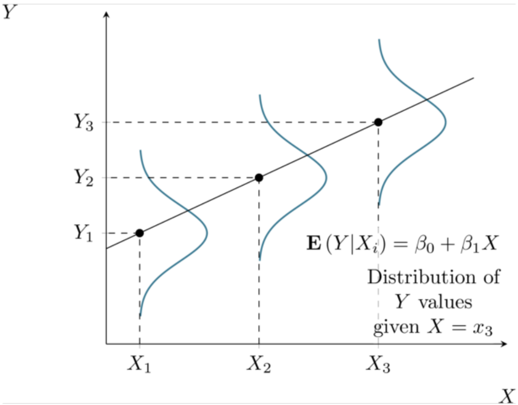
$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}; \quad n > p$$

- The OLS estimate of regression coefficient: $\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.

Illustration of linear regression model

Least squares fit

- If we collect height-weight data for the population, there will be many "weight" data points corresponding to height = x1 = 160. Similarly there will be many "weight" data points corresponding to height = x2 = 170; x3 = 180 etc.
- The estimated regression line connects the average weights for each of the height sub populations.



Interpretation of the Model Parameters

- Each slope β represents the change in the mean response, $E(y)$, per unit increase in the associated predictor variable when all the other predictors are held constant.
- For example, β_1 represents the change in the mean response, $E(y)$, per unit increase in x_1 when $x_2, x_3, ..., x_p$ are held constant.
- The intercept term, β_0, represents the mean response, E(y), when all the predictors $x_1, x_2, ..., x_p$ are all zero (which may or may not have any practical meaning).

Significance Testing of Each Variable

- The standard error of an estimator SE($\hat{\beta}_j$) reflects how it varies under repeated sampling
- To determine whether a variable x_j is a useful predictor variable in this model, we could test

  H_0: There is no relationship between x_j and y
  H_1: There is some relationship between x_j and y
- Mathematically, this corresponds to testing

$$H_0 : \beta_j = 0$$

versus

$$H_0 : \beta_j \neq 0$$

Significance Testing of Each Variable

- T-test:

$$t_0 = \frac{\hat{\beta}_j - 0}{\mathrm{SE}(\hat{\beta}_j)}$$

  follows a t distribution with degree of freedom $n - (p + 1)$
- p-value = $P(|t| > t_0)$
- When we cannot reject the null hypothesis above ($p - value > \alpha$), we should say that we do not need variable x_j in the model given that other variables will remain in the model.
- confidence interval $\beta_j \pm t_\alpha \mathrm{SE}(\hat{\beta}_j)$

Assessing the Overall Accuracy of the Model

Sum of squares and analysis of variance

- Consider the identity $y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$
- Squaring both sides and summing over all n observations we get

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$$

- Note: $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})e_i = \sum_{i=1}^{n}\hat{y}_i e_i - \bar{y}\sum_{i=1}^{n}e_i = 0$
- Hence we have

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

SST (Total SS) = SSR (Regression SS) + SSE (Error SS)

Assessing the Overall Accuracy of the Model

How well does the model fit the data?
Coefficient of determination R^2

- $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2, \ SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \ SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- In the absence of any information, our best guess for a new y is $\bar{y}$.
- We expect the estimated values $\hat{y}_i$'s to be close to the data points and hence they are far from $\bar{y}$ making the proportion SSR/SST large.
- Define, Coefficient of determination = $R^2 = SSR/SST = 1 - SSE/SST$
- R^2 is a measure of assessing the strength of linear relationship. It represents the proportion of variation explained by the regressor x.
- Since $0 \leq SSE \leq SST$, it follows $0 \leq R^2 \leq 1$.
- $R^2 = 1$ implies all data points fall perfectly on regression line.
- $R^2 = 0$ implies estimated regression line is perfectly horizontal.

Assessing the Overall Accuracy of the Model

R^2 and adjusted R^2

- Both R^2 and R^2_{Adj} can be used for assessing the overall adequecy of the model.
- $R^2 = SSR/SST = 1 - SSE/SST$
- R^2 never decreases when a regressor is added to a model, regardless of the value of the contribution of that variable.
- $R^2_{Adj} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$
- R^2_{Adj} will only increase on adding a variable to the model if the addition of the variable reduces the MSE.
- R^2_{Adj} penalizes the model for adding the terms that are not helpful, thus making it useful in evaluating and comparing candidate regression models.

Assessing the Overall Accuracy of the Model

Is at least one of the predictors $x_1, ..., x_p$ useful in predicting the response?
Hypotheses:

$$H_0 : \beta_1 = \beta_2.....\beta_p = 0$$

$$H_1 : \text{at least one } \beta_j \text{ is not zero}$$

Analysis of Variance (ANOVA) table

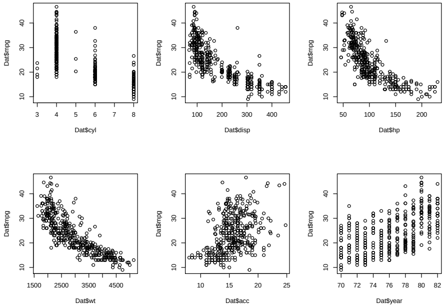
Source of Variation	DF	SS	MS	F	P-value
Regression	p	$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	MSR = SSR/p	$F_0 = \frac{MSR}{MSE}$	$P(F > F_0)$
Residual (Error)	n-p-1	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	MSE = SSE/(n-p-1)		
Total	n-1	$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$			

- If $P - value < \alpha$, reject H_0, there is at least one variable is significant to the model

Auto MPG Data Set

Auto MPG Data Set is available at UCI machine learning repository (http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data)

1. mpg: miles per gallon
2. cyl: cylinders
3. disp: displacement
4. hp: horsepower
5. wt: weight
6. acc: acceleration
7. year: model year



Building a MLR model: R output

```
autompg <- read.csv("/datasets/autompg.csv")
Dat <- autompg

## Include the functions required for data partitioning
source("/Lecture Notes/myfunctions.R")


## Scatter plot ##
par(mfrow = c(2, 3))
plot(Dat$cyl, Dat$mpg)
plot(Dat$disp, Dat$mpg)
plot(Dat$hp, Dat$mpg)
plot(Dat$wt, Dat$mpg)
plot(Dat$acc, Dat$mpg)
plot(Dat$year, Dat$mpg)


RNGkind (sample.kind = "Rounding")
set.seed(0) ## set seed so that you get same partition each time
p2 <- partition.2(Dat, 0.7) ## creating 70:30 partition
training.data <- p2$data.train
test.data <- p2$data.test
```

Building a MLR model: R output

```
> ## Fit MLR model
> mlr <- lm(mpg ~ ., data = training.data)
>
> ## Inference on model parameters ##
> summary(mlr)

Call:
lm(formula = mpg ~ ., data = training.data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6579 -2.6075 -0.1924  2.0558 14.1537

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.622e+01  6.064e+00  -2.675  0.00793 **
cyl         -3.906e-02  4.307e-01  -0.091  0.92781
disp        -1.478e-03  9.610e-03  -0.154  0.87791
hp          -1.111e-03  1.852e-02  -0.060  0.95219
wt          -6.398e-03  9.063e-04  -7.060 1.45e-11 ***
acc          9.274e-02  1.287e-01   0.720  0.47187
year         7.608e-01  6.635e-02  11.466  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.589 on 266 degrees of freedom
Multiple R-squared:  0.7961,Adjusted R-squared:  0.7915
F-statistic: 173.1 on 6 and 266 DF,  p-value: < 2.2e-16
```

Interpreting the R output

Interpretation of regression coefficients

- $\hat{y}_i$: predicted response (fitted value) for experimental unit i
- $\hat{\beta}_0 = -16.2226$; $\hat{\beta}_1 = -0.0391$ $\hat{\beta}_2 = -0.0015$ $\hat{\beta}_3 = -0.0011$ $\hat{\beta}_4 = -0.0064$ $\hat{\beta}_5 = 0.0927$ $\hat{\beta}_6 = 0.7608$
- Fitted MLR model for automobile data:
  $\hat{mpg}_i = -16.2226 - 0.0391 cyl - 0.0015 disp - 0.0011 hp - 0.0064 wt + 0.0927 acc + 0.7608 year$

Interpreting the R output

Interpretation of regression coefficients β_0

- The intercept $\hat{\beta}_0$ indicates the expected value of y at $x_1 = \cdots = x_p = 0$.
- In the current context $\hat{\beta}_0 = -16.2226$ *i.e.* a car with cyl = disp = hp = wt = acc = year = 0 is expected to have -16.2226 mpg. This definitely does not make any sense.
- It happens because we extrapolated beyond the **scope** of the model (range of x values). $\hat{\beta}_0$ is meaningful only if the scope of the model includes $x_j = 0$.

Interpreting the R output

Interpretation of regression coefficients β_j for $j = 1, ..., p$

- The slope $\hat{\beta}_j$ indicates the change in the expected value of y per unit increase in x_j.
- $\hat{\beta}_1 = -0.0391$ implies that we expect the mean mpg to decrease by 0.0391 unit for every 1 unit increase in cyl when the other regressors are held constant.
- $\hat{\beta}_6 = 0.7608$ implies that we expect the mean mpg to increase by 0.7608 unit for every 1 unit increase in year when the other regressors are held constant.

Interpreting the R output

Hypothesis testing concerning β_j

$$H_0 : \beta_j = 0 \ vs. \ H_1 : \beta_j \neq 0$$

This can be answered in two ways:

- We can check the **p-value** for each regression coefficient in the summary output. If p-value is less than pre-specified significance level $\alpha = 0.05$, reject H_0 and conclude that there is a linear association between the response variable and the regressor.
- We can obtain the **confidence interval**.

Interpreting the R output

Hypothesis testing concerning β_j

- The linear relationship between *mpg* and *cyl* is not significant.
- The linear relationship between *mpg* and *disp* is not significant.
- The linear relationship between *mpg* and *hp* is not significant.
- The linear relationship between *mpg* and *wt* is significant.
- The linear relationship between *mpg* and *acc* is not significant.
- The linear relationship between *mpg* and *year* is significant.

Interpreting the R output

$100 \times (1 - \alpha)\%$ confidence interval for β_j

```
> ## 95% Confidence interval for model parameters ##
> confint(mlr, level = 0.95)
                  2.5 %       97.5 %
(Intercept) -28.162563913 -4.282686883
cyl          -0.887048941  0.808929368
disp         -0.020399608  0.017444218
hp           -0.037579319  0.035356433
wt           -0.008182367 -0.004613652
acc          -0.160705050  0.346189261
year          0.630160690  0.891443872
```

- If we take 100 different samples, and compute 95% confidence interval for each of them, then 95 of those 100 intervals will contain the true parameter value
- If the confidence interval contains 0, then we conclude that there is not enough evidence for a linear relationship at $100 \times (1 - \alpha)\%$ confidence level.

Interpreting the R output

Note: This does not necessarily mean that there is no relationship between x_j and y.

- Maybe the relationship is not linear.
- Maybe there is a linear relationship, but we failed to reject H_0 when it is actually false (Type II error).

Interpreting the R output

Assessing the Overall Accuracy of the Model

```
Residual standard error: 3.589 on 266 degrees of freedom
Multiple R-squared:  0.7961,Adjusted R-squared:  0.7915
F-statistic: 173.1 on 6 and 266 DF,  p-value: < 2.2e-16
```

- Residual standard error: $MSE = \frac{SSE}{n-(p+1)} = 3.589$
- Coefficient of determination: $R^2 = 0.7961$ and $R_{Adj} = 0.7915$
  79.15% of variation in y can be explained by the regression equation
- The P-value for the F test is almost zero.
  We reject the hypothesis that there is no correlation between $\boldsymbol{x}$ and y at all

Multicollinearity

- Multicollinearity exists when two or more of the predictors in a regression model are moderately or highly correlated.
- Recall: $\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$. The near linear dependence between regressors may result in a singular $\boldsymbol{X}' \boldsymbol{X}$.
- Multicollinearity is usually measured by Variance Inflation Factor (VIF) (R function *vif* from *car* package)
- For the j^{th} regressor, $VIF_j = \frac{1}{1-R_j^2}$, where R_j^2 is the coefficient of determination obtained from regressing x_j on other regressor variables.
- Generally, VIF value $>$ 4 is a matter of concern (VIF $>$ 10 is definitely a matter of concern)
- Multicollinearity may result in regression coefficients having wrong sign.

VIF: R output

```
> ## Check multicollinearity
> library(car)
> vif(mlr)
      cyl      disp        hp        wt       acc      year
11.107550 19.497625  9.446665 11.948585  2.672559  1.239140
```

- Model suffers from potential multicollinearity.
- Check correlation.
- *cyl* variable is highly correlated with *disp*, *hp* and *wt*.
- Let us remove *cyl* from model.

```
> ## Check correlation
> cor(training.data)
            mpg        cyl       disp         hp         wt        acc       year
mpg   1.0000000 -0.7685389 -0.8008870 -0.7797178 -0.8217858  0.3954530  0.5728735
cyl  -0.7685389  1.0000000  0.9537763  0.8520819  0.8941795 -0.4915418 -0.3330027
disp -0.8008870  0.9537763  1.0000000  0.8879373  0.9335071 -0.5076415 -0.3459772
hp   -0.7797178  0.8520819  0.8879373  1.0000000  0.8671756 -0.6613591 -0.4054616
wt   -0.8217858  0.8941795  0.9335071  0.8671756  1.0000000 -0.3754948 -0.2954267
acc   0.3954530 -0.4915418 -0.5076415 -0.6613591 -0.3754948  1.0000000  0.2605957
year  0.5728735 -0.3330027 -0.3459772 -0.4054616 -0.2954267  0.2605957  1.0000000
> ## Remove highly correlated variable
> mlr1 <- lm(mpg ~ disp+hp+wt+acc+year, data = training.data)
> vif(mlr1)
     disp        hp        wt       acc      year
10.386853  9.446660 11.929373  2.670104  1.239025
```

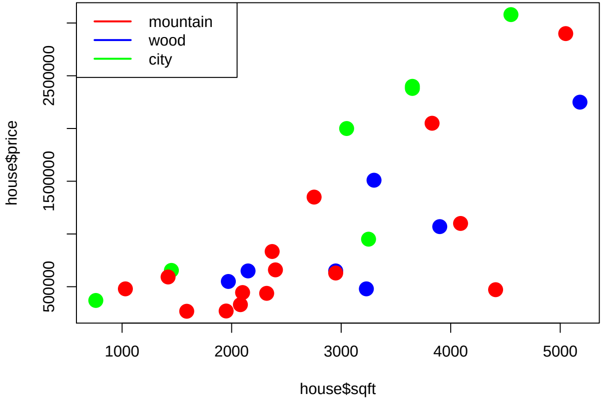
Prediction using MLR model

- Predict the mean *mpg* for a car with
  $cyl = 6, disp = 200, hp = 140, wt = 3220, acc = 12, year = 80$

```
> ## Given value for x ##
> x0 <- data.frame(cyl=6, disp=200, hp=140, wt=3220, acc = 12, year=80)
>
> ## Forecasting mean response for a given value of x
> predict(mlr, x0)
       1
24.46737
```

MLR with categorical regressors

- Dataset: houseprice
- Response: price
- Regressors: sqft, view (levels: wood, city, mountain)
- For a categorical variable with d levels, create $d - 1$ dummy variables



MLR model: R output

```
mlr <- lm(price ~ sqft + factor(view) , data=house)
> summary(mlr)

Call:
lm(formula = price ~ sqft + factor(view), data = house)

Residuals:
     Min       1Q   Median       3Q      Max
-1302814  -219402     1065   312577   782859

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           135088.1   300363.8   0.450  0.65677
sqft                     534.9       81.1   6.595 6.56e-07 ***
factor(view)mountain -719122.4   225878.1  -3.184  0.00387 **
factor(view)wood     -845262.3   264322.1  -3.198  0.00374 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 491900 on 25 degrees of freedom
Multiple R-squared:  0.697,Adjusted R-squared:  0.6606
F-statistic: 19.17 on 3 and 25 DF,  p-value: 1.15e-06
```

- $view = city$: $\widehat{price_i} = 135088.1 + 534.9 sqft$
- $view = mountain$: $\widehat{price_i} = 135088.1 + 534.9 sqft - 719122.4$
- $view = wood$: $\widehat{price_i} = 135088.1 + 534.9 sqft - 845262.3$

MLR model: R output

```
mlr <- lm(price ~ sqft + factor(view) , data=house)
> summary(mlr)

Call:
lm(formula = price ~ sqft + factor(view), data = house)

Residuals:
     Min       1Q   Median       3Q      Max
-1302814  -219402     1065   312577   782859

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           135088.1   300363.8   0.450  0.65677
sqft                     534.9       81.1   6.595 6.56e-07 ***
factor(view)mountain -719122.4   225878.1  -3.184  0.00387 **
factor(view)wood     -845262.3   264322.1  -3.198  0.00374 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 491900 on 25 degrees of freedom
Multiple R-squared:  0.697,Adjusted R-squared:  0.6606
F-statistic: 19.17 on 3 and 25 DF,  p-value: 1.15e-06
```

- For every unit increase in *sqft* the mean *price* is expected to increase by 534.9 unit when *view* is held constant.
- If *view* is changed from *city* to *mountain* the mean *price* is expected to decrease by 719122.4 unit when *sqft* is held constant.
- If *view* is changed from *city* to *wood* the mean *price* is expected to decrease by 845262.3 unit when *sqft* is held constant.

MLR model: R output

- to change the baseline category of the *view* variable, use the *relevel*() function

```
mlr <- lm(price ~ sqft + relevel(view, ref = "wood") , data=house)
> summary(mlr)

Call:
lm(formula = price ~ sqft + relevel(view, ref = "wood"), data = house)

Residuals:
     Min       1Q   Median       3Q      Max
-1302814  -219402     1065   312577   782859

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                       -710174.2   321904.8  -2.206  0.03678 *
sqft                                  534.9       81.1   6.595 6.56e-07 ***
relevel(view, ref = "wood")city    845262.3   264322.1   3.198  0.00374 **
relevel(view, ref = "wood")mountain 126139.9   229561.5   0.549  0.58755
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 491900 on 25 degrees of freedom
Multiple R-squared:  0.697,Adjusted R-squared:  0.6606
F-statistic: 19.17 on 3 and 25 DF,  p-value: 1.15e-06
```

- do not drop "relevel(view, ref = "wood")mountain" because all dummies together constitutes the categorical regressor "view"