NBA 4920/6921 Lecture 3

Linear Regression Part 1

Murat Unal

Johnson Graduate School of Management

09/07/2021

Agenda

- Quiz 2
- Linear regression
- Inference
- Model performance
- Interpreting output
- Modeling interactions
- Qualitative predictors

Load/install the following packages.

Download the Advertising data and load it into R.

Read in the Credit and Auto data from the ISLR package

```
rm(list=ls())
options(digits = 3, scipen = 999)
library(tidyverse)
library(ISLR)
library(cowplot)
library(ggcorrplot)
library(stargazer)
library(corrr)
library(lmtest)
library(sandwich)
library(MASS)
library(car)
library(jtools)
data <- read.csv("Advertising.csv")
credit <- ISLR::Credit
auto <- ISLR::Auto
```

Linear regression

Linear regression is a simple parametric approach to supervised learning

It assumes the relationship between the outcome Y and the inputs $X = X_1, X_2, \cdots, X_p$ is linear

It's conceptually simple and easy to implement

The model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

We obtain estimates for the coefficients $\beta_0, \beta_1, \cdots, \beta_p$ by minimizing the **Residual Sum of Squares** (RSS)

$$
\begin{aligned}
RSS &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2
\end{aligned}
$$

The values $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ are the least squares coefficient estimates

Inference

The standard error of an estimator reflects how it varies under repeated sampling.

Standard errors can be used to compute confidence intervals.

A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.

It has the form

$$\hat{\beta}_j \pm 2 \cdot SE(\hat{\beta}_j)$$

Standard errors can also be used to perform hypothesis tests on the coefficients.

In regression setting we test the **null hypothesis** of

H_0 : The coefficient $\hat{\beta}_j$ has no effect on Y, i.e
$\hat{\beta}_j = 0$ *versus*

the **alternative hypothesis**

H_A : The coefficient $\hat{\beta}_j$ has some effect on Y i.e
$\hat{\beta}_j \neq 0$

To test the null hypothesis, we compute a $t - stat$, given by

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

This will have a t-distribution with $n - 2$ degrees of freedom, assuming $\hat{\beta}_j = 0$

The *p-value* is the probability of observing any value equal or greater than $|t|$, we reject H_0 if $p \leq 0.05$

We can also test for **any** association between the predictors and the response, i.e.

we can answer *if at least one predictor is useful.*

The **null hypothesis** now becomes

H_0 : All coefficients have no effect on Y, i.e
$$\hat{\beta}_1 = \hat{\beta}_2 = \cdots = \hat{\beta}_p = 0$$

versus the **alternative hypothesis**

H_A : At least one $\hat{\beta}_j$ is non-zero

To test the null hypothesis, we compute the $F - stat$, given by

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{n,n-p-1}$$

$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, $TSS = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$ is the **Total Sum of Squares**

TSS represents the RSS using the mean of the outcome only, i.e. a model with no predictors

The $F - stat$ will be much larger than 1 if there is any relationship and we reject H_0 if $p \leq 0.05$

Comparing nested models

Two regression models are called nested if one contains all the predictors of the other, and some additional predictors.
For example, the model in two independent variables,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

is nested within the model in four independent variables.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

How to choose between them?

If the larger model has just one more predictor than the smaller model, you could just test the significance of the one additional coefficient, using the t-statistic.

When the models differ by $q > 1$ added predictors, you cannot compare them using t-statistics.

The conventional test is based on comparing the residual sums of squares for the two models

Since a model with additional predictors will always reduce the residual sum of squares, we ask whether this reduction is statistically meaningful.

Let RSS_f, RSS_q be the residual sum of squares from a large and small models, respectively.

Then the $F - statistic$ becomes:

$$F = \frac{(RSS_q - RSS_f)/q}{RSS_f/(n-p-1)}$$

In R we can use the `anova()` function to implement this comparison.

Model performance

How does our linear model fit the data? We want to quantify it.

Using RSS and TSS we compute **Residual Standard Error** (RSE) and **R-squared**(R^2)

$$RSE = \sqrt{\frac{RSS}{n-p-1}}, \quad R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

RSE is the average amount the response deviates from the regression line.

It measures the **lack of fit** of the model in absolute terms, i.e units of Y

R^2 represents the **fraction of Variance** explained by our model, independent of the scale of Y

R^2 close to 1 indicates that a large proportion of the variability in Y has been explained by our model

Adding more variables to the model always increases R^2, whereas RSE can increase or decrease

As such, we need to be careful about overfitting, especially if we aim for prediction

R^2 provides no protection against overfitting, quite opposite - **encourages** it

Application

```
str(data)
```

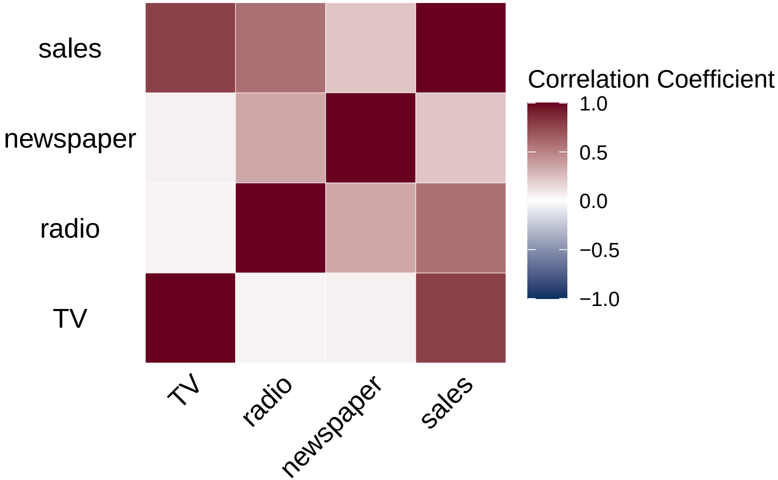
```
'data.frame':   200 obs. of  4 variables:
 $ TV       : num  230.1 44.5 17.2 151.5 180.8 ...
 $ radio    : num  37.8 39.3 45.9 41.3 10.8 48.9 32.8 19.6
 $ newspaper: num  69.2 45.1 69.3 58.5 58.4 75 23.5 11.6 1
 $ sales    : num  22.1 10.4 9.3 18.5 12.9 7.2 11.8 13.2 4.
```

Lets' check the correlations and visualize the relationships between `Sales` and advertising in different media:

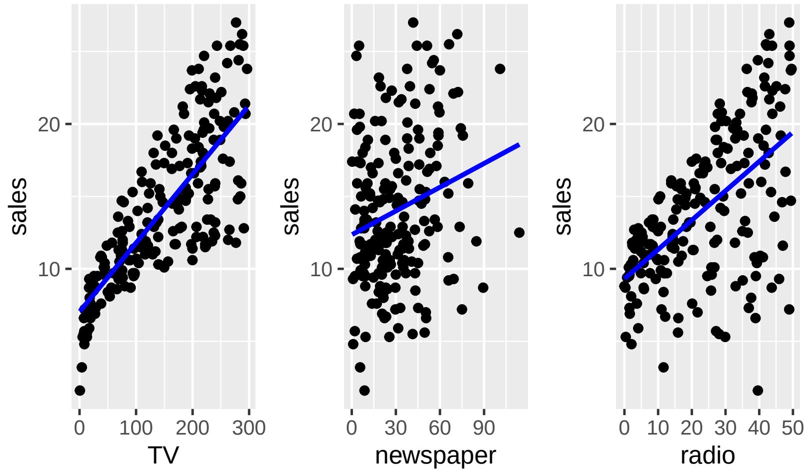
```
corr <- cor(data)
corr
```

```
              TV  radio newspaper sales
TV        1.0000 0.0548    0.0566 0.782
radio     0.0548 1.0000    0.3541 0.576
newspaper 0.0566 0.3541    1.0000 0.228
sales     0.7822 0.5762    0.2283 1.000
```

```
ggcorrplot(corr, type = "full",lab = FALSE,
       legend.title = "Correlation Coefficient",
       colors = c("#053061", "white", "#67001f"),
       ggtheme = ggplot2::theme_void,
       outline.col = "white")
```



```r
p1 <- ggplot(data,mapping = aes(x =TV,y=sales)) +
      geom_point() +
      geom_smooth(method = "lm", formula = y~x,
                  se=FALSE,colour = "blue")
p2 <- ggplot(data,mapping = aes(x =newspaper,y=sales)) +
      geom_point() +
      geom_smooth(method = "lm", formula = y~x,
                  se=FALSE,colour = "blue")
p3 <- ggplot(data,mapping = aes(x =radio,y=sales)) +
      geom_point() +
      geom_smooth(method = "lm", formula = y~x,
                  se=FALSE,colour = "blue")
plot_grid(p1,p2,p3, ncol = 3)
```



sales
20
10
0
100
200
300
TV
sales
20
10
0
30
60
90
newspaper
sales
20
10
0
10
20
30
40
50
radio

Use the `lm()` function to run a regression and `summary()` or `summ()` to get the output.

```
lm1 <- lm(sales~TV+radio+newspaper, data = data)
summary(lm1)
summ(lm1)
```

Interpreting model output

We interpret β_j as the average effect on Y of a one unit increase in X_j, holding all other predictors **fixed**.

```
MODEL FIT:
F(3,196) = 570.271, p = 0.000
R² = 0.897
Adj. R² = 0.896

Standard errors: OLS
---------------------------------------------------
                      Est.    S.E.    t val.      p
----------------- -------- ------- -------- -------
(Intercept)          2.939   0.312    9.422   0.000
TV                   0.046   0.001   32.809   0.000
radio                0.189   0.009   21.893   0.000
newspaper           -0.001   0.006   -0.177   0.860
---------------------------------------------------
```

Didn't we see a positive relationship between `newspaper` and `sales`?

Why do we get a negative coefficient for the effect of `newspaper`?

Correlations among input variables can be problematic as changing one variable will simultaneously change the correlated variables.

Newspaper and radio ads are correlated to each other as well as to sales, leaving one out inflates the effect of the included.

```r
corr <- cor(data)
corr
```

```
              TV  radio newspaper sales
TV        1.0000 0.0548    0.0566 0.782
radio     0.0548 1.0000    0.3541 0.576
newspaper 0.0566 0.3541    1.0000 0.228
sales     0.7822 0.5762    0.2283 1.000
```

Let's see this by running the regressions separately:

```r
lm.TV <- lm(sales~TV, data = data)
lm.radio <- lm(sales~radio, data = data)
lm.newspaper <- lm(sales~newspaper, data = data)
```

```r
summ(lm.TV,model.info=FALSE,model.fit=FALSE,digits=3)
```

```
Standard errors: OLS
--------------------------------------------------
                      Est.    S.E.   t val.       p
----------------- ------- ------- -------- -------
(Intercept)          7.033   0.458   15.360   0.000
TV                   0.048   0.003   17.668   0.000
--------------------------------------------------
```

The estimate for `TV` didn't change much.

```r
summ(lm.radio,model.info=FALSE,model.fit=FALSE,digits=3)
```

```
Standard errors: OLS
------------------------------------------------
                     Est.    S.E.   t val.      p
----------------- ------- ------- -------- -------
(Intercept)         9.312   0.563   16.542   0.000
radio               0.202   0.020    9.921   0.000
------------------------------------------------
```

```
summm(lm.newspaper,model.info=FALSE,model.fit=FALSE,digits=3
```

```
Standard errors: OLS
-------------------------------------------------------
                       Est.     S.E.    t val.        p
----------------- -------- ------- -------- -------
(Intercept)          12.351    0.621    19.876    0.000
newspaper             0.055    0.017     3.300    0.001
-------------------------------------------------------
```

But the estimates for both `radio` and `newspaper` changed significantly. More importantly `newspaper` now has a positive effect, because leaving out the correlated variable inflates the effect of the included.

This is the **Omitted Variable Bias** in effect and is the reason we refrain from making **causal claims** in regression settings with observational data

Mainly because we can never conclusively argue that we have accounted for all variables that might be correlated simultanesously with the dependent and one or more of the independent variables

What other variables do you think we might be missing here?

Let's check if the model is useful overall

```r
summary(lm1)$fstat
```

```
value numdf dendf
  570     3   196
```

```r
round(pf(summary(lm1)$fstat[1], summary(lm1)$fstat[2],
   summary(lm1)$fstat[3], lower.tail = FALSE),3)
```

```
value
    0
```

F-stat is large and the associated *p-val* is ≤ 0.01

Notice, the same info is being produced after calling the summ() function.

```
summ(lm1,model.info=FALSE,digits=3)
```

```
MODEL FIT:
F(3,196) = 570.271, p = 0.000
R² = 0.897
Adj. R² = 0.896

Standard errors: OLS
---------------------------------------------------
                      Est.    S.E.   t val.       p
----------------- -------- ------- -------- -------
(Intercept)          2.939   0.312    9.422   0.000
TV                   0.046   0.001   32.809   0.000
radio                0.189   0.009   21.893   0.000
newspaper           -0.001   0.006   -0.177   0.860
---------------------------------------------------
```

```
summary(lm1)$r.squared
```

```
[1] 0.897
```

The model explains 90% of the variability in `sales`

```r
summary(lm1)$sigma
```

```
[1] 1.69
```

On average predicted `sales` values will deviate by 1.69 units or dollars

So, yes, we conclude that the model is useful overall in explaining sales as a function of the advertising expenditure in different media

Let's compare the model to one that has only `TV` as predictor.

```
anova(lm.TV,lm1)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
198	2103	NA	NA	NA	NA
196	557	2	1546	272	0

The *F-stat* is large and the associated *p-val* is ≤ 0.01, so using the larger model is justified.

Modeling interactions

We can enrich the linear model by including interactions if we expect that the effect of one variable might not be constant but depend on the magnitude of another variable.

In the advertising model, for example, the advertsing expenditure on `radio` can actually increase the effectiveness of `TV` advertising.

If this is the case then the slope term for `TV` will not be constant and should increase as `radio` increases.

We can test this idea with the following model:

$$\begin{aligned} Sales &= \beta_0 + \beta_1 TV + \beta_2 radio + \beta_3 radio \times TV + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 radio) \times TV + \beta_2 radio + \epsilon \end{aligned}$$

```
lm.interact <- lm(sales~TV*radio,data = data)
```

```r
summ(lm.interact, model.info=FALSE,digits=4)
```

```
MODEL FIT:
F(3,196) = 1963.0569, p = 0.0000
R² = 0.9678
Adj. R² = 0.9673

Standard errors: OLS
---------------------------------------------------
                     Est.      S.E.     t val.        p
----------------- -------- -------- --------- --------
(Intercept)        6.7502    0.2479   27.2328   0.0000
TV                 0.0191    0.0015   12.6990   0.0000
radio              0.0289    0.0089    3.2408   0.0014
TV:radio           0.0011    0.0001   20.7266   0.0000
---------------------------------------------------
```

There's strong evidence in favor of rejecting
$H_0 : \hat{\beta}_3 = 0$

Which suggest that the effect of `TV` ad spending on `sales` depends on the level of `radio` ad spending

An increase in `TV` ad spending of $1000 is associated with increased `sales` of

$$(\hat{\beta}_1 + \hat{\beta}_3 \times \mathit{radio}) \times 1000 = 19 + 1.1 \times \mathit{radio} \text{ units}$$

What about the impact of `radio` ad spending?

What about the impact of `radio` ad spending?

An increase in `radio` ad spending of \$1000 is associated with increased `sales` of

$$(\hat{\beta}_2 + \hat{\beta}_3 \times TV) \times 1000 = 29 + 1.1 \times TV \ \text{ units}$$

Let's visualize this interaction effect.

First pick three values for `radio` from its distribution:

```
summary(data$radio)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0    10.0    22.9    23.3    36.5    49.6 
```

Let's pick the first quartile, the mean, and the third quartile: 10,23,36.

Now, we want to obtain new predictions at three levels for radio and all the TV data using the model we just estimated.

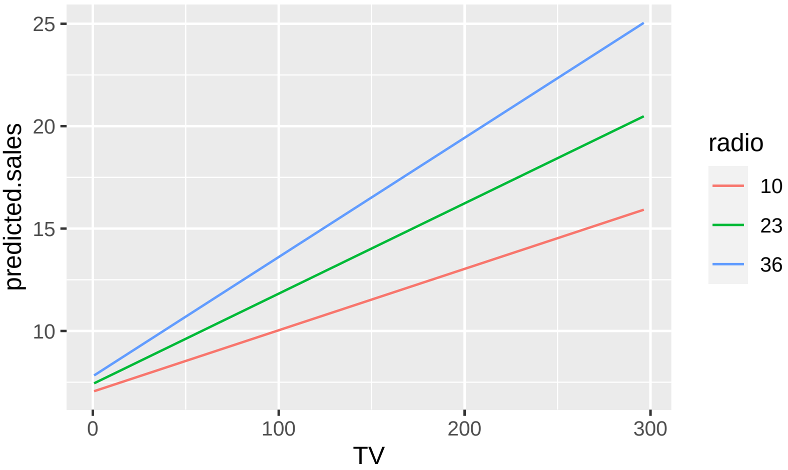
Create a new data:

```
new.data = data.frame(TV = rep(data$TV,3),
                      radio=c(rep(10,200),
                      rep(23,200),rep(36,200)))
```

And use the `predict()` function for this new data:

```
new.data$predicted.sales <- predict(lm.interact,
                                    newdata = new.data)
```

```r
new.data$radio = factor(new.data$radio)
ggplot(new.data, mapping=aes(x=TV,y=predicted.sales,
                                        colour=radio))+
  geom_line()
```



The hierarchy principle

If your model includes interactions follow the hierarchy principle: include the main variables as well, even if their associated *p-values* are not statistically significant

Qualitative predictors

In order to include qualitative/categorical/factor variables such as sex, marital status, race into the model we need to define new binary variables

For example if we want to include a race variable in our model we define two new variables:

$$x_{i1} = \begin{cases} 1 \text{ if } i\text{th person is Asian} \\ 0 \text{ if } i\text{th person is not Asian} \end{cases}$$

$$x_{i2} = \begin{cases} 1 \text{ if } i\text{th person is Caucasian} \\ 0 \text{ if } i\text{th person is not Caucasian} \end{cases}$$

The model then becomes

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i \text{ if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i \text{ if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i \text{ if } i\text{th person is African American} \end{cases}$$

The level with no dummy variable is the **baseline**

Now β_0 can be interpreted as the average credit card balance for African Americans,

β_1 can be interpreted as the difference in the average balance between the Asian and African American categories, and

β_2 can be interpreted as the difference in the average balance between the Caucasian and African Americans

Let's apply this in the credit data.

```
str(credit)
```

```
'data.frame':	400 obs. of  12 variables:
 $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Income   : num  14.9 106 104.6 148.9 55.9 ...
 $ Limit    : int  3606 6645 7075 9504 4897 8047 3388 7114
 $ Rating   : int  283 483 514 681 357 569 259 512 266 491
 $ Cards    : int  2 3 4 3 2 4 2 2 5 3 ...
 $ Age      : int  34 82 71 36 68 77 37 87 66 41 ...
 $ Education: int  11 15 11 11 16 10 12 9 13 19 ...
 $ Gender   : Factor w/ 2 levels " Male","Female": 1 2 1 2
 $ Student  : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1
 $ Married  : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 1 1
 $ Ethnicity: Factor w/ 3 levels "African American",..: 3 2
 $ Balance  : int  333 903 580 964 331 1151 203 872 279 135
```

`Ethnicity` is a factor variable with three levels.

Let's create two dummy variables for Asian and Caucasian and run the regression of `Balance` against these new dummies.

```r
credit$Asian = ifelse(credit$Ethnicity=="Asian",1,0)
credit$Caucasian = ifelse(credit$Ethnicity=="Caucasian",
                                                      1,0)
```

```r
lm.dummy <- lm(Balance~Asian + Caucasian, data=credit)
summ(lm.dummy,model.info=FALSE,digits=3)
```

```
MODEL FIT:
F(2,397) = 0.043, p = 0.957
R² = 0.000
Adj. R² = -0.005

Standard errors: OLS
-----------------------------------------------------
                       Est.      S.E.    t val.     p
----------------- --------- -------- -------- -------
(Intercept)         531.000   46.319   11.464   0.000
Asian               -18.686   65.021   -0.287   0.774
Caucasian           -12.503   56.681   -0.221   0.826
-----------------------------------------------------
```

We could also just use the `factor()` function in R without creating additional dummies.

```
lm.dummy <- lm(Balance~factor(Ethnicity), data=credit)
summ(lm.dummy,model.info=FALSE,digits=3)
```

```
MODEL FIT:
F(2,397) = 0.043, p = 0.957
R² = 0.000
Adj. R² = -0.005

Standard errors: OLS
-------------------------------------------------------
                                     Est.      S.E.    t va
--------------------------------- --------- -------- ------
(Intercept)                         531.000   46.319   11.4
factor(Ethnicity)Asian              -18.686   65.021   -0.2
factor(Ethnicity)Caucasian          -12.503   56.681   -0.2
-------------------------------------------------------
```

How do you interpret this result?

We see that the estimated balance for the baseline, African American, is \$531.00.

It is estimated that the Asian category will have \$18.69 less debt than the African American category,

and that the Caucasian category will have \$12.50 less debt than the African American category.

However, the p-values associated with the coefficient estimates for the two dummy variables are very large, suggesting no statistical evidence of a real difference in credit card balance between the ethnicities

Exercise

1. Run a regression of `Sales` against main variables and all possible interactions. Use the syntax `lm(sales~.^2,data)`.
2. Compare this full interaction model to the one that has only the main variables. Is the interaction model justified?
3. Use the first quartile, the mean, and the third quartile of `newspaper`, fix `radio` at its mean, and plot the interaction effect of `TV*newspaper`.

```r
lm.full.int <- lm(sales~.^2,data)
summ(lm.full.int, model.info=FALSE,
                  model.fit=FALSE,digits =3)
```

```
Standard errors: OLS
------------------------------------------------------
                           Est.    S.E.   t val.       p
--------------------- -------- ------- -------- -------
(Intercept)               6.460   0.318   20.342   0.000
TV                        0.020   0.002   12.633   0.000
radio                     0.023   0.011    2.009   0.046
newspaper                 0.017   0.010    1.691   0.092
TV:radio                  0.001   0.000   19.930   0.000
TV:newspaper             -0.000   0.000   -2.227   0.027
radio:newspaper          -0.000   0.000   -0.464   0.643
------------------------------------------------------
```

```r
anova(lm1,lm.full.int)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
196	557	NA	NA	NA	NA
193	170	3	387	146	0

```r
summary(data$newspaper)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.3    12.8    25.8    30.6    45.1   114.0
```

```r
new.data = data.frame(TV=rep(data$TV,3),
                      newspaper=c(rep(12.8,200),
                      rep(30.6,200),rep(45.1,200)),
                      radio=rep(mean(data$radio),600))
```

```r
new.data$predicted.sales <- predict(lm.full.int,
                                    newdata = new.data)
```

```
new.data$newspaper = factor(new.data$newspaper)
ggplot(new.data, mapping=aes(x=TV,y=predicted.sales,
                                       colour=newspaper))+
  geom_line()
```